5-2013

The Development of Introductory Statistics Students' Informal Inferential Reasoning and Its Relationship to Formal Inferential Reasoning

Bridgette Lynn Jacob
Syracuse University

Abstract

The difficulties introductory statistics students have with formal statistical inference are well known in the field of statistics education. "Informal" statistical inference has been studied as a means to introduce inferential reasoning well before and without the formalities of formal statistical inference. This mixed methods study investigated the development of introductory statistics students' informal inferential reasoning and its relationship to their formal inferential reasoning. A pre/posttest was administered to 136 students enrolled in introductory statistics classes taught in their secondary schools. Four task-based interviews were also conducted with seven pairs of those students.

With probabilistic reasoning essential for formal statistical inference, students' informal inferential reasoning related to sampling and estimating probabilities did improve significantly. Additionally, the strong informal inferential reasoners, those strong at the beginning and remaining so at the end of the study, and the 14 students who took part in the task-based interviews demonstrated strong formal inferential reasoning at the end of their course. However, when the interviewees drew informal inferences based on a sampling distribution, the majority of them sought to reduce variability by taking several samples. This study provided insight into why students were not using the probabilities associated with the normality of the sampling distribution when drawing an informal inference and how this then impacted their formal inferential reasoning. Implications for practice and suggestions for further research are included.

The Development of Introductory Statistics Students' Informal Inferential Reasoning and Its Relationship to Formal Inferential Reasoning

By

Bridgette L. Jacob
B.A. Niagara University, 1983
M.S. Syracuse University, 2000

DISSERTATION

Submitted in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in Mathematics Education
in the Graduate School of Syracuse University

May 2013

Acknowledgements

I would like to start by thanking Helen Doerr for sharing her valuable research expertise with me during my doctoral studies and this dissertation study. For the many times when I could not see my way clearly through the research process, she knew how to help me frame the work I had done and still needed to do to move forward. I always left our meetings feeling certain that this was within my reach and invigorated to pick up where I had left off. I truly appreciate your time, your guidance, and your friendship.

Thank you to Hyune-Ju Kim and James Bellini for their work as members of my committee. Your perspectives and expertise were invaluable to me throughout the dissertation process. To the teachers who opened their classrooms to me and supported the work I did with their students, needless to say, I could not have completed this study without your help. Thank you so very much.

I would also like to thank Helen Doerr and Joanna Masingila for providing a wonderfully supportive atmosphere to grow professionally at Syracuse University. You both lead by example and have had a significant impact on my life as a teacher of mathematics and now as a researcher in mathematics education. To my fellow doctoral students whom I have had the pleasure of spending time with in research seminar and at conferences, I thank you for the camaraderie and the valuable feedback you have offered during this journey.

To all of my family and friends who have provided continual encouragement over the past four years, thank you. I especially thank my husband and my two children who have been with me every step of the way. I simply could not have done this without your love and support.

Table of Contents

List of Tables and Figures

Tables

Figures

Chapter One - Introduction

The past four decades have seen unparalleled growth in the field of statistics, especially in its applications, with statistics courses now found in almost every post-secondary school. The inclusion of probability and statistics as one of the content standards in the National Council of Teachers of Mathematics standards document (NCTM, 1989) made statistics part of the mathematics curriculum in grades K-12 in the United States. Since the first administration of the Advanced Placement Statistics examination in 1997, the number of high school students taking the examination has grown from 7,600 to over 100,000 students from around the world. This has triggered increased interest in how statistics is taught and how students learn statistics among statistics educators. While much progress has been made in statistics education, it is common to find that students, even after instruction in an introductory statistics course, are lacking the ability to properly apply inferential reasoning. Applying this reasoning incorporates an understanding of the basic concepts of descriptive statistics, basic probability rules, the sampling distribution, and how these work together in analyzing data. The analysis of data known as inferential reasoning, drawing conclusions and making decisions based on samples of data, lies at the heart of any statistical investigation.

Aims of this Research

The analysis of data is referred to as "formal" statistical inference and consists of estimating population parameters with confidence intervals and testing conjectures about population parameters with hypothesis tests. The study of formal statistical inference in an introductory statistics course generally follows the study of descriptive statistics, probability, and the sampling distribution. The difficulties students have in bringing these concepts together to conduct formal statistical inference are well known in the field of statistics education (Castro

Sotos, Vanhoof, Noortgate, & Onghena, 2009; Garfield & Ben-Zvi, 2008; Tversky & Kahneman, 1974; Konold, 1989; Konold et al., 2011). Most recently, "informal" statistical inference has been studied as a means to bridge the gap between descriptive statistics and formal statistical inference. In “informal” statistical inference, students draw conclusions about populations based on data without the formalities of constructing a confidence interval or conducting a hypothesis test. Instead, the data are carefully reviewed to understand the main features (e.g., center and spread) to determine what evidence this data might provide about the population.

The statistical reasoning students employ while working on informal statistical inference tasks is termed “informal inferential reasoning.” A thorough understanding of how informal inferential reasoning develops may help to build the bridge to formal inferential reasoning, the reasoning that ties together the concepts of descriptive statistics, probability, and the sampling distribution for formal statistical inference. The focus of this study involved investigating the development of students’ informal inferential reasoning as well as the strength of the relationship between their informal and formal inferential reasoning. The research questions I investigated in this study were:

1) For students enrolled in an introductory statistics class:

    a) Does their informal inferential reasoning develop?

    b) If their informal inferential reasoning develops, what are the characteristics of this informal inferential reasoning as it develops?

2) What is the relationship between students' informal inferential reasoning and their formal inferential reasoning?

Rationale for this Research

Formal statistical inference involves conducting a hypothesis test or constructing a confidence interval with data from a sample, then drawing appropriate conclusions about a population. Although students are taught the procedures of formal statistical inference, many of which they can remember and demonstrate, this does not necessarily mean they are able to draw appropriate conclusions about populations based on data or fully comprehend the implications of those conclusions. A thorough understanding of the underlying concepts and how they work together is required to make informed decisions with appropriate data analysis.

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (2010) was developed by statisticians and statistics educators, endorsed and funded by the American Statistical Association (ASA). Based on research in statistics education, the report emphasizes conceptual understanding of the procedures necessary for statistical analysis. Among the goals put forth in the report are the following regarding statistical inference.

> Students should understand the basic ideas of statistical inference, including:
>
> - the concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the idea of standard error);
> - the concept of statistical significance, including significance levels and *p*-values; and
> - the concept of confidence interval, including the interpretation of confidence level and margin of error. (GAISE College Report, 2010, p. 12)

Research in statistics education has shown that understanding the underlying concepts taught throughout an introductory statistics course often does not provide students with the support to achieve the conceptual understanding necessary for formal statistical inference. What

may appear to be the next step in learning statistics to those who are seasoned in the subject, is actually a large chasm for many introductory statistics students.

Recent research efforts have focused on "informal" statistical inference to understand how students begin to reason about data they encounter (Ben-Zvi, 2006; Pfannkuch, 2006; Pratt, Johnston-Wilder, Ainley, & Mason, 2008). Informal inferential reasoning is characterized by students drawing conclusions about populations based on data before the procedures of formal inference are introduced. While several related definitions of informal inferential reasoning have been proposed, they share an underlying theme of inferring characteristics of a population from a sample data while recognizing the uncertainty that exists. This focus on informal inferential reasoning as students take part in informal statistical inference tasks has been under investigation for the last decade. As an example, one type of problem used to study students' informal inferential reasoning is comparing two distributions of data displayed in boxplots or histograms (Ben-Zvi, 2004; Pfannkuch, 2006; Watson, 2002; Watson & Moritz, 1999). In these problems, students are asked to draw conclusions about populations well before they have been introduced to formal statistical inference. While researchers are building definitions of informal inferential reasoning and frameworks for its development, exactly how informal inferential reasoning develops and how students demonstrate it is still under investigation. This research study was designed to add to the understanding of that development by investigating how students demonstrated informal inferential reasoning as they worked through carefully designed tasks and to examine the nature and strength of the relationship between their informal and formal inferential reasoning.

Theoretical Frameworks

Two theoretical frameworks were used for this study. First, in my investigation of how students' informal inferential reasoning developed, I used the principles essential to informal statistical inference proposed by Makar and Rubin (2009). Initial concepts that Makar and Rubin saw as essential for informal inferential reasoning were:

- notion of uncertainty and variability articulated through language that broke from the mathematical convention of claims of certainty;
- reliance on the concept of aggregate (as opposed to individual points) through the use of generalizations about the group;
- acknowledgement of a mechanism or tendency that extended beyond the data at hand; and
- evidence for reasoning based on purposeful use of data. (p. 85)

From these initial concepts, Makar and Rubin then developed three principles essential to informal statistical inference:

> (1) *generalization*, including predictions, parameter estimates, and conclusions, that extend *beyond describing the given data*; (2) the use of data *as evidence* for those generalizations; and (3) employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn. (p. 85)

Using these principles, I sought to document the development of students' informal inferential reasoning as they were engaged in informal statistical inference tasks. Evidence of their informal inferential reasoning was determined by the extent to which they (1) made an inference based on

the data, (2) used the data as evidence for their inference, and (3) used probabilistic language to indicate a level of certainty in their inference.

The second theoretical framework concerned the types of tasks used to investigate students' informal inferential reasoning. In a preliminary study, I employed a task design framework developed by Zieffler, Garfield, delMas, and Reading (2008). These authors proposed three categories of tasks to conduct research on informal inferential reasoning:

- estimate and draw a graph of a population based on a sample;
- compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled; and
- judge which of two competing models or statements is more likely to be true. (p. 47)

Based on the results of the preliminary study, I chose to modify these task categories slightly by turning to the research on students' difficulties and misconceptions with statistical inference. The tasks in this study were related to those proposed by Zieffler et al. (2008) and offered a clear link to students' informal inferential reasoning as it developed during their study of introductory statistics. The first task in this study corresponded to the second type in the Zieffler et al. framework and had students comparing distributions of data to make an informal inference. This type of task drew on the research on difficulties students encounter with the concepts of variation and distribution in descriptive statistics (Bakker & Gravemeijer, 2004; Ben-Zvi, 2004; Kelly & Watson, 2002; Makar & Confrey, 2005; Reading & Shaughnessy, 2000; Shaughnessy, Canada, & Ciancetta, 2003; Watson, 2002; Watson & Moritz, 1999). Students' misconceptions of sampling and basic probability were the basis for the second task in my study. With this task, students made inferences about unknown population probabilities based on their own sampling. The third task in this study was related to the third category proposed by Zieffler et al. which

stemmed from students' difficulties in understanding the sampling distribution. In this task, students used the sampling distribution to make an informal inference based on a sample of data. This sequence of tasks provided a means to gain insight into students' informal inferential reasoning throughout their study of introductory statistics.

The review of literature in Chapter 2 examines research on students' difficulties with variation and distribution, probability, and the sampling distribution in support of these chosen tasks while situating this study overall in the current body of research.

Chapter 2 - Related Literature

In this review of literature, I return to the GAISE guidelines to examine the research findings that address the three basic concepts of the sampling distribution, statistical significance, and confidence intervals that support the goal of an understanding of statistical inference. An understanding of the sampling distribution is required to determine if data from a sample can be viewed as likely or unlikely under the assumed conditions of a population. Research has been conducted on students' difficulties with the sampling distribution and its relationship to statistical inference. Determining the likelihood or significance of sample data and interpreting the results of a confidence interval requires knowledge of basic probability theory. Researchers have found that intuitive probabilistic reasoning can lead to biases which translate into incorrect conclusions regarding statistical significance and confidence intervals. Researchers trace students' difficulties with sampling distributions, statistical significance, and confidence intervals back to the basic concepts of variation and distribution. Therefore, I will also review research on students' understandings of variation and distribution. The most recent area of research examines what researchers have called "informal inferential reasoning" which involves making inferences based on data but without the computational formalities. Researchers have studied students' work as they compare distributions of data, for example, to understand how questions and activities can promote this informal inferential reasoning for a deeper understanding of formal statistical inference. Frameworks have also been proposed for the study of informal inferential reasoning. To conclude, I will discuss the implications of this research for my intended study.

The Sampling Distribution

A conceptual understanding of the sampling distribution is required for formal statistical inference. This includes an understanding of the shape of the sampling distribution and the effect

of sample size on the sampling distribution. It should be noted that for most introductory statistics courses, the focus is on approximately normal sampling distributions. Well, Pollatsek, and Boyce (1990) conducted a series of four experiments designed to investigate how students understood the effects of sample size on the variability of the mean and the sample mean. The first three experiments consisted of multiple choice questions administered to 114 introductory psychology students asking them to consider the mean of a smaller sample in comparison to the mean of a larger sample in a problem context. Given the value of the population mean, students were to determine which sample mean would most likely be closer to the population mean (termed "center" version questions) and which would deviate more from the population mean (termed "tail" version questions). Results revealed that students performed significantly better on the "center" version questions than the "tail" version questions ($p \leq .05$). The researchers conjectured that

> tail problems focus subjects on the degree of discrepancy between the populations and the sample averages and, for some subjects, this apparently elicits an inappropriate heuristic about extreme scores, namely, that extreme scores are more likely to occur in large samples (which is true) and that, therefore, the averages of large samples will be more variable (which is not true). (p. 310)

In the fourth experiment, instruction on random sampling and the sampling distribution was provided to the students in a structured interview format that included questions equivalent to those administered in the first three experiments. Results showed that even after instruction on the sampling distribution in which students observed the decrease in variability of sample averages with samples of size 10 compared to the population, students could not make the distinction of less variability in the averages of larger samples.

Saldanha and Thompson (2002) designed a study to examine students developing ideas about repeated sampling and the sampling distribution while they participated in instruction on the topic. Their study was based on the knowledge gained from studies similar to that of Well, Pollatsek, and Boyce (1990) and others that found students focus on the statistics of samples, the sample mean, for example, rather than on how these statistics are distributed. Saldanha and Thompson state that "sampling has not been characterized in the literature as a scheme of interrelated ideas entailing repeated random selection, variability, and distribution" (p. 258). Therefore, instruction for their teaching experiment stressed two themes: "1) the random selection process can be repeated under similar conditions, and 2) judgments about sampling outcomes can be made on the basis of relative frequency patterns that emerge in collections of outcomes of similar samples" (p.259). Their study took place in a non-AP statistics course with 27 eleventh and twelfth grade students over nine class sessions. To gain information on students' understandings, classroom sessions were videotaped, students' written work was evaluated, and individual interviews were conducted after the experiment.

Students worked with computer simulations in which they took many samples from a variety of populations with known parameters. As evidenced through students' work and classroom discussions, Saldanha and Thompson conjectured that the instructional activities helped some students develop "a multi-tiered scheme of conceptual operations centered around the images of repeatedly sampling from a population, recording a statistic, and tracking the accumulation of statistics as they distribute themselves along a range of possibilities" (p. 261). However, the majority of the students still compared a single sample statistic to the population parameter rather than to the sampling distribution of all such statistics when asked to determine if it was unusual. This type of reasoning has the potential to create problems in formal statistical

inference, particularly in hypothesis testing when the likelihood of a sample statistic is determined in relation to the sampling distribution. Saldanha and Thompson called this an "additive" image of sample in contrast to a "multiplicative" conception of sample where the distinction between the population, the individual samples taken from the population, and the distribution of many such samples is made. They claim that the "multiplicative" conception of sample will help students to understand why statisticians can be confident in inferring about the population based on data from a single sample.

In a research project spanning seven years, Chance, delMas, and Garfield used interactive software designed to assist students in making these distinctions between the population, samples, and the sampling distribution. The project consisted of five studies focused on the difficulties college students in introductory statistics classes encountered when learning about sampling distributions. The three researchers collaborated on all activities, lessons, and assessments administering them in three settings described as a small private university, a College of Education, and a Developmental Education College. In their first study, delMas, Garfield, and Chance (1998) assessed students' understanding of sampling distributions after using simulation software that allowed them to change the shape of a population distribution and the sample size used to construct the sampling distribution while viewing how these affected the sampling distribution. Students were then given several paper-and-pencil problems each displaying a population distribution and five possible related sampling distributions. They were asked, for example, to select the sampling distribution of 500 samples of sample size four and also of sample size 25. Pretests and posttests revealed that students were not developing a clear concept of the sampling distribution in terms of the impact that the population distribution and sample size had on the sampling distribution. To help with this, for the second study delMas,

Garfield, and Chance (1999) asked students to respond to the questions first, or make predictions, and then to test their predictions using the software. This resulted in significant increases on their posttest results ($p < .05$).

In their third study, Chance, delMas, and Garfield (2004) used data collected in the first two studies and their reflections to specify common misconceptions as follows:

- believe sampling distribution should look like the population (for sample size n > 1).
- think sampling distribution should look more like the population as the sample size increases (generalizes expectations for a single sample of observed values to a sampling distribution).
- predict that sampling distributions for small and large sample sizes have the same variability.
- believe sampling distributions for large samples have more variability.
- do not understand that a sampling distribution is a distribution of sample statistics.
- confuse one sample (real data) with all possible samples (in distribution) or potential samples.
- pay attention to the wrong things, for example, heights of histogram bars.
- think the mean of a positive skewed distribution will be greater than the mean of the sampling distribution for samples taken from this population. (p. 302)

This provided insight into why students were experiencing difficulties with the sampling distribution.

The fourth and fifth studies consisted of interviews with students in introductory statistics classes at the undergraduate and graduate levels as they interacted with the same simulation software and similar paper-and-pencil problems. Following the fourth study, the authors built a

model describing the developmental stages that students progress through as they come to understand the sampling distribution. However, in attempting to validate these stages as part of the fifth study, they realized that students could not be placed in one developmental stage. Instead, based on the interviews from the fifth study, Chance, delMas, and Garfield (2004) proposed the following dimensions of statistical reasoning behavior:

- *Fluency* – how well the student understands and uses appropriate terminology, concepts, and procedures
- *Rules* – the degree to which a student identifies and uses a formal rule to make predictions and explanations
- *Consistency* – the presence or absence of contradictory statements
- *Integration* – the extent to which ideas, concepts, and procedures are connected
- *Equilibrium* – the degree to which a student is aware of any inconsistencies or contradictions in his or her statements and predictions
- *Confidence* – the degree of certainty in choices and statements (p. 309)

These studies together demonstrate the difficulties students encounter when introduced to the sampling distribution. Many of these difficulties surround their misunderstandings of the effects of sample size and, thus, the variability of the sampling distribution. In addition, once students have an initial perception of the sampling distribution, placing the sampling distribution in relation to the population and individual samples is problematic. The difficulties students have with the sampling distribution plague them further as they attempt to navigate statistical inference, namely the concept of statistical significance and the interpretation of confidence intervals.

Statistical Significance and Confidence Intervals

Many of the difficulties students have with statistical significance are due to their probabilistic reasoning. Work done by psychologists revealed that individuals' intuitive probabilistic reasoning about events was not always based on appropriate probability theory. Psychologists Tversky and Kahneman (1974) described three heuristics people used to make judgments when faced with uncertainty. The first was "representativeness" which is used to determine if an event (A) comes from a process (B). "For example, when A is highly representative of B, the probability that A originates from B is judged to be high" (p. 1124). Use of this reasoning ignores the existing probability that event A would occur. For example, in tossing a fair coin six times, tossing three heads and three tails would be thought to have a greater probability than tossing six heads. This heuristic can also lead to what psychologists and educators call the "law of small numbers" in which even small samples are judged to be representative of a population if they portray the assumed features of the population. For example, in an experiment to determine if a particular coin is fair, tossing two heads and two tails in four tosses may be considered enough evidence to conclude the coin is fair.

Another heuristic for determining the probability of an event described by Tversky and Kahneman (1974) is "availability" in which the probability of an event is determined "by the ease with which instances or occurrences can be brought to mind" (p. 1127). Probabilities calculated in this manner are based on an individual's own experiences which may be in contrast to the actual probability of an event. For example, a person may believe the probability of having an accident on the way to work is zero if this has never happened.

The third heuristic described by Tversky and Kahneman (1974) is termed "adjustment" which occurs when the probability of an event is determined by an adjustment from an initial

related probability. For example, even if the probability of an event is high, the probability of this event occurring in succession will be lower. Often the probability of a succession of events is overestimated based on an adjustment of the probability of one such occurrence of the event. For example, the probability of tossing a head with a fair coin is one-half. The probability of tossing a sequence of six heads may be thought to be close to one-half when it is actually considerably less.

These three heuristics for making judgments or calculating probabilities have been demonstrated by students new to the study of probability and statistics as well as by individuals trained in probability theory. These intuitive methods for reasoning are resistant to change since they can be correct in several day-to-day situations. Judgments that students make when inferring about a population based on sample data can be influenced by these intuitive probabilistic reasonings, contradicting judgments that should be made with proper probability theory and thorough statistical analysis. Use of these three heuristics, "representativeness", "availability" and "adjustment", reveals a lack of consideration and understanding of variability, distribution, and the important concept that probabilities emerge in the long-run.

Konold (1989) believed that these judgment heuristics were incomplete. He added the "outcome approach" as another method by which people make decisions. Konold interviewed 16 introductory psychology students, as they solved problems involving uncertainty. With the results from these interviews, he identified two features of the "outcome approach" to reasoning: "(a) the tendency to interpret questions about the possibility of an outcome as requests to predict the outcome of a *single trial* and (b) the reliance on *causal* as opposed to stochastic explanations of outcome occurrence and variability" (p. 65). In subsequent interviews with 12 of these students five months later, Konold found the results to be consistent when using related

problems. Further validity of the "outcome approach" to reasoning occurred with the ability to predict the responses to problems not used in the first interview that would draw out this same type of reasoning.

Konold (1989) identified two types of statements supporting students' tendency to answer probability questions with the single trial feature of the outcome approach: a yes/no response about one outcome (e.g., a 70% chance of rain on a particular day meant that it would rain); and a right/wrong response when evaluating a prediction (e.g., if it did not rain when the chance was 70%, the prediction was incorrect). These types of reasonings can impact student interpretations of statistical significance in hypothesis testing and confidence intervals. For example, statistically significant results do not indicate that the null hypothesis is false and a confidence interval may not contain the population parameter.

More recent work by Konold et al. (2011) examined an eighth grade student's understanding of the relationship between experimental and theoretical probabilities. Two of the researchers, Konold and Kazak, conducted an interview with this student, Erin, after she demonstrated a rather advanced understanding of this relationship. They were holding a series of after-school workshops with a focus on data analysis and probability in which they were testing new activities with the *TinkerPlots* software. This investigation took place during a one-hour session in which the authors gave this student a series of tasks to complete. The first was the Spinner Problem using *TinkerPlots* in which Erin was to predict the probability of getting green based on simulated trials. The second task, also using *TinkerPlots*, was the Die Problem which was similar to the first in that Erin was to predict the probability of getting a "2" with a fair die. For the third task, the Bone Problem, Erin was given an irregular shaped bone with six sides and asked which side would most likely land on top when the bone was rolled.

During the interview, Erin demonstrated more confidence in the experimental probabilities due to the fact that there would be variability and the exact theoretical probability was not likely to occur in any trial. Two main observations of Erin's perceptions of probability emerged. The first was that she lacked the understanding these trials were estimates of a "true" probability that is unknown. Additionally, in the Bone Problem, when a "true" probability was nearly impossible to theorize, Erin did not believe that conducting more trials would give her a better estimate. This lead to the second observation which was that Erin did not have a clear understanding of the Law of Large Numbers. Both of these perceptions could prove inhibitive for an understanding of statistical inference.

In an effort to develop an assessment to capture students' understandings of statistical concepts including their probabilistic reasoning at the completion of an introductory statistics course, delMas, Garfield, Ooms, and Chance (2007) developed the Comprehensive Assessment of Outcomes in Statistics (CAOS). Over a period of three years, the authors piloted and revised a series of multiple-choice questions. The CAOS 4 assessment exceeded acceptable levels for internal consistency with a Cronbach's alpha coefficient of 0.82 and was used with a sample of 763 introductory statistics students enrolled in 20 different two-year and four-year colleges and universities from 14 different states in the United States. Results of the pretest and posttest analysis revealed statistically significant gains ($p < .001$) in students' ability to recognize the desirability of small *p*-values in studies, that the *p*-value is not the probability that a treatment is effective, and in making a correct interpretation when the null hypothesis is rejected. Similar gains were shown in students' abilities to correctly interpret a confidence interval as well as recognizing that the confidence level (95% for example) does not represent the percentage of population data values or sample means falling in the interval. However, overall students scored

low on both the pretest and the posttest on questions in which they had to correctly interpret the *p*-value. Additionally, they misinterpreted the *p*-value as the probability that a treatment was not effective and a confidence level as the percentage of sample data falling in the interval. For both *p*-values and confidence intervals, students demonstrated that they could recognize a correct interpretation but also indicated that incorrect interpretations were valid. This indicates the lack of a thorough understanding of both of these concepts.

In a review of research on reasoning about statistical inference, Garfield and Ben-Zvi (2008) found that survey studies administered to undergraduate and graduate students, instructors, and scientists "have identified persistent misuses, misinterpretations, and common difficulties people have in understanding of inference, statistical estimation, significance tests, and *p*-values" (p. 265). They summarize common misconceptions that students have regarding *p*-values and confidence intervals.

Misconceptions about *p*-values

- A *p*-value is the probability that the null hypothesis is true.
- A *p*-value is the probability that the null hypothesis is false.
- A small *p*-value means the results have significance (statistical and practical significance are not distinguished).
- A *p*-value indicates the size of an effect (e.g., strong evidence means big effect).
- A large *p*-value means the null hypothesis is true, or provides evidence to support the null hypothesis.
- If the *p*-value is small enough, the null hypothesis must be false.

Misconceptions about Confidence Intervals

- There is a 95% chance the confidence interval includes the sample mean.

- There is a 95% chance the population mean will be between the two values (upper and lower limits).
- 95% of the data are included in the confidence interval.
- A wider confidence interval means less confidence.
- A narrower confidence interval is always better (regardless of confidence level). (p. 270)

More recent research on students' misconceptions has been conducted by Castro Sotos, Vanhoof, Noortgate, and Onghena (2009). They studied misconceptions in statistical inference, specifically in hypothesis testing, *p*-values, and significance level. These researchers administered a questionnaire to 144 undergraduate students with three of five multiple-choice questions pertaining to hypothesis tests. Students showed a lack of understanding of exactly what the result of a hypothesis test means. Twenty percent of the students responded that a hypothesis test proves or disproves the null hypothesis. They were also still prone to confusing the probabilistic meanings of the *p*-value and the significance level. For example, 21% of the students identified the *p*-value as the probability of the null hypothesis and 16% identified it as the probability of incorrectly rejecting the null hypothesis. In addition, 17% identified the significance level as the probability of the null hypothesis, and 17% answered that 1 - α (the significance level) was the probability of rejecting the null hypothesis.

Many researchers trace the difficulties students have with sampling distributions, statistical significance, and confidence intervals back to the basic concepts of variation and distribution. Thus, a body of research exists on these two fundamental concepts.

Variation and Distribution

Wild and Pfannkuch (1999) established a framework for statistical thinking that identified "thinking patterns involved in problem solving, strategies for problem solving, and the integration of statistical elements within the problem solving" (p. 224). Their four-dimensional framework came about as a result of their review of relevant literature and interviews with statisticians and statistics students to reveal their statistical reasoning processes. In their Dimension 2: Types of Thinking, variation is seen as fundamental to statistical thinking with the following components: noticing and acknowledging variation; measuring and modeling for the purposes of prediction, explanation, or control; explaining and dealing with variation; and investigative strategies including the use of randomization. Variation, a key concept in statistical thinking, involves coming to terms with the implicit uncertainty contained in data. It is the ever-present variation that creates the need for "sophisticated statistical methods to filter out any messages in data" (p. 236). These statistical methods include, for example, hypothesis testing when determining whether a sample of data is likely to occur under the null hypothesis is based on the variation that is expected in the sampling distribution.

A sequence of studies surrounding the Gumball Task or Lollies Task, as it is called in Australia, was explored after the 1996 National Assessment of Educational Progress (NAEP) results were published in America. In the Lollies Task, a bowl was filled with certain percentages of different colored candies. Students were given these percentages and asked to determine how many candies of one color they would expect in a handful of 10 candies. The question was worded to elicit an exact response from the students. Reading and Shaughnessy (2000) examined this task to determine what version would motivate students to begin to think about variation. These researchers found that a version of the question asking students in grades

four through 12 to give a possible sample of candies from the jar was most effective in getting students to think about variation. They were asked to conjecture about the number of candies if six people drew 10 each. The candies were returned and mixed thoroughly after each draw. As students progressed in age they demonstrated increased ability to describe the sampling situation. However, students overall were lacking in their ability to give reasons for their responses.

This task was further explored by Kelly and Watson (2002) who added an interview component to probe students' thinking. These students in grades three through nine were also given the opportunity to conduct the experiment themselves. Results showed that students demonstrated intuitions about both center and spread. Students' explanations of their responses showed an increased ability to reason about the distributions and proportions of candies up to grade seven. There was a decrease in performance level for the grade nine students which the authors attributed to the fact that these students were selected from classes considered to be of average ability. This was in contrast to students chosen from grades three through seven who were considered to be of average to high ability.

Shaughnessy, Canada, and Ciancetta (2003) further revised the Gumball Task for a study with middle school students on prediction tasks. The revised tasks included a Sampling Task (Gumball Task with candies), the Dice Task of repeated rolls of a die, and the Spinner Task of repeated outcomes of a spinner. Results showed that students' predictions of outcomes for these tasks included variation. In fact, 70% of the student responses were considered reasonable in terms of spread or variation for the Sampling Task. That decreased to 48% for the Spinner Task and to 30% for the Dice Task. These authors conjecture that this low percentage for rolls of a die is most likely due to previous instruction involving probabilities of such outcomes without examining the variability that appears with repeated trials.

This development of activities surrounding distribution and variation gave more insight into what students knew and could articulate about both center and variation that exists in any sampling situation. While variability occurs within any sampling situation and data set, another layer of variability occurs between data sets. Watson and Moritz (1999) investigated how comparing data sets allows students to develop skills for later inferential reasoning. In their comparisons, they used data sets of equal size and data sets of different sizes. Eighty-eight students from two Australian schools in grades three through nine were asked to compare graphical displays of these data sets. These researchers observed three hierarchical levels of student responses. The first level of response was one in which only one aspect of the graphical representation of the data was recognized such as the center. A response at the second level occurred when a student recognized more than one aspect of the distribution such as where the peak of the distribution occurred and that it was skewed. At this level, these different aspects were not viewed as working together to describe the distribution. Therefore, this level response often caused a conflict for students that they were not able to reason through. Students responding at the third and highest level were able to see several aspects of the data including distribution and variation. For these students, "the whole has a coherent structure and meaning" (p. 149). It is at this level that students will be best poised to make inferences about populations based on sample data.

Later, Watson (2002) expanded this research by introducing a component of cognitive conflict. Individual interviews with 20 students each from third, sixth, and ninth grade were conducted. The student responses were recorded before and after the introduction of other students' remarks about comparing the distributions. In response to the distributions of data sets of equal size, 57% of the students initially responding incorrectly improved their responses.

When comparing distributions of data sets of different sizes, 30% improved their responses. Watson concluded that the use of cognitive conflict benefited students in understanding statistical inference.

Ben-Zvi (2004) focused on the development of students' understanding of variability from a local perspective, within a data set, to a global perspective of describing variability between data sets. He conducted a study of two seventh grade students working together with technology on an exploratory data analysis task. He found seven distinct stages in the students' understanding of variation as it progressed from a local to a global view. These stages were:

1. On what to focus: Beginning from irrelevant and local information;
2. How to describe variability informally in raw data;
3. How to formulate a statistical hypothesis that accounts for variability;
4. How to account for variability when comparing groups using frequency tables;
5. How to use center and spread measures to compare groups;
6. How to model variability informally through handling outlying values; and
7. How to notice and distinguish the variability within and between the distributions in a graph. (p. 48)

Ben-Zvi recommended that students work together on comparison tasks to develop their understanding of variability within and between data sets.

These studies show how variation is intertwined with distributional concepts such as center and shape as students recognize and articulate the variation that exists in a single data set as well as between data sets. A clear understanding of both is essential for success in inferential reasoning tasks. This was seen in Saldanha and Thompson's (2002) exploration of high school students' understanding of sampling and sampling distributions with simulation data. While

students were able to reason about the variability among samples, this reasoning did not necessarily translate to the variation that existed in sample statistics. This then "obstructed their ability to imagine how sample proportions might distribute themselves around the underlying population proportion" (p. 264). Students' limited view of variation kept them from a distributional view of sample proportions. Without this distributional view, the underlying premise of hypothesis testing becomes extremely difficult for students to understand. Students' abilities to relate a distribution of sample data to the population distribution is crucial for understanding formal inferential reasoning.

Researchers have also explored how students express their understanding of variation and distribution. Conducting 12 to 15 lessons in four seventh-grade classrooms in the Netherlands, Bakker and Gravemeijer (2004) found that students had some basic ideas about distributions of data which they expressed with their own informal language. They used terms such as "bump" or "hill" to describe the distribution of the data and reasoned how this might become higher or wider with more data. Using technology, these students analyzed data in graphs and constructed graphs, without the use of formal terms such as mean or spread.

Makar and Confrey (2005) investigated how teachers constructed the meaning behind measures of variation like standard deviation. They examined the discourse of preservice math and science teachers, listening to the nonstandard language used when comparing two distributions. Makar and Confrey found that even though students "may be taught the concept and formula for standard deviation and can even use this term as part of class discourse, this does not imply that they are 'seeing' variation in what is being measured" (p. 31). Knowing the procedures for statistical calculations does not ensure that students grasp the depth of the significance behind them. Students' nonstandard language becomes an important component of

understanding what notions they have when comparing distributions of data. In their analysis, Makar and Confrey found that these preservice teachers did have some understanding of the concepts of variation or spread and of distribution such as center and shape. In fact, their nonstandard language often encompassed both notions at the same time. This suggests that learners could develop their informal inferential reasoning before they are expected to comprehend the formalities of statistical inference.

Informal Inferential Reasoning

Recently, informal inferential reasoning has been explored as a means of building the bridge between an understanding of variation and distribution in descriptive statistics and formal statistical inference. Studies examining the impact of carefully designed questions asking students to draw conclusions based on graphical displays of data, some using statistical software, have begun to identify the components of informal inferential reasoning and how it might be developed in students.

Pfannkuch (2006) defined informal inferential reasoning as the "drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data" (p. 1). Pfannkuch worked with a year 11 teacher as she explored informal inferential reasoning with her class of 15-year-old students. In a teaching episode comparing the number of text messages sent by users of two different cell phone companies, students were to decide which phone company users sent the most text messages in a month. The data was displayed graphically in two boxplots. Although the teacher referred to the underlying distributions of data sets verbally, this was not prevalent in her written conclusions when comparing the data sets. The students' communications revealed that they lacked the ability to see the data as referring to the underlying distributions. Pfannkuch argues for developing communication in the classroom

to improve inferential reasoning as well as giving students the chance to explore how samples from differing distributions behave.

Ben-Zvi (2006) conducted a study with fifth-graders using the statistical software *TinkerPlots* so students could develop and analyze displays of data. He argued that even though beginning statistical ideas such as organizing data and exploratory data analysis were now being introduced at the primary level, the informal inferential reasoning component was often lacking. Ben-Zvi stated that "Deriving logical conclusions from data - whether formally or informally - is accompanied by the need to provide persuasive arguments based on data analysis" (p. 2). Results showed that the fifth-graders considered aggregate views of data, understood the importance of larger samples of data, and accounted for variability. Of equal importance was their ability to express and justify their claims.

Pratt, Johnston-Wilder, Ainley, and Mason (2008) examined local and global thinking when students reasoned informally. They used a working definition proposed by Makar and Rubin (2007): "We consider informal inferential reasoning of statistics in broad terms to be the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected" (p.1). Pratt and his colleagues described a classroom situation which may inhibit students' ability to reason inferentially. They described this as the difference between situations where the data represents the entire population and those in which the data is a sample from the population. Students will have difficulties looking beyond the data at hand if they believe that the data they have is all that exists. Using software designed by Pratt for the study, students had the ability to add to an existing sample or generate a new sample. In either situation, students tended to focus on the changes in subsequent displays of the data. At times when they did express a global understanding by referring to the stability found when considering all of the

samples, they were still frustrated by the fluctuations they saw in the individual samples. This suggested that an important aspect of informal inference is in finding the invariance that is present even among all of the local changes.

Statistics education research over the past two decades has focused on students' formal inferential reasoning. This is an important area since students' understanding of how to draw appropriate conclusions about populations from sample data lies at the heart of the purpose and usefulness of statistics in the many fields where statistical analysis is needed. Researchers have progressed from studying students' misconceptions and difficulties when analyzing data to examining students' understanding of distributions of data and the variations that exist in the data. Students' understanding of distribution and variation are thought to impact these misconceptions and difficulties. More recently, researchers have begun to study informal inferential reasoning which incorporates the concepts of distribution and variation in drawing conclusions about populations without the formalities of confidence intervals and hypothesis tests. Researchers have proposed several frameworks for the study of informal inferential reasoning.

Frameworks for the Study of Informal Inferential Reasoning

To facilitate the study of informal inferential reasoning, researchers have developed frameworks which attempt to partition informal inferential reasoning into its components. One of the first such frameworks was based on the qualitative analysis of teaching episodes of a grade 11 teacher by Pfannkuch (2006). As a result of this analysis, Pfannkuch proposed a teaching model for informal inferential reasoning that might be used when teaching students how to compare boxplots. This model included eight elements of reasoning that were "non-hierarchical,

interdependent but distinguishable" (p. 33) and two moderating elements of reasoning contained in each of the eight elements of reasoning:

Elements of Reasoning	
1. Hypothesis generation	Compares and reasons about the group trend.
2. Summary:	Compares equivalent five-number summary points. Compares non-equivalent five-number summary points.
3. Shift:	Compares one box plot in relation to the other box plot and refers to comparative shift.
4. Signal:	Compares the overlap of the central 50% of the data.
5. Spread:	Compares and refers to type of spread/densities locally and globally within and between box plots.
6. Sampling:	Considers sample size, the comparison if another sample was taken, the population on which to make an inference.
7. Explanatory:	Understands context of data, considers whether findings make sense, considers alternative explanations for the findings.
8. Individual case:	Considers possible outliers, compares individual cases.

Moderating Elements of Reasoning

9. Evaluative:	Evidence described, assessed on its strength, weighed up.
10. Referent:	Group label, data measure, statistical measure, data attribution, data plot distribution, contextual and statistical knowledge. (p. 33)

Pfannkuch stated that the evaluative and referent elements of reasoning "act as anchors for weighing the evidence and for interpreting an abstract box plot representation respectively" (p. 42). The elements of reasoning in this teaching model were proposed to improve teachers' and students' abilities to make informal inferences when comparing distributions.

Watson (2008) adapted the eight elements of reasoning in Pfannkuch's framework, but without the moderating elements of reasoning. This framework was used in a study of teaching what she terms “beginning” inference to a class of 15 seventh-grade students and was also used to assess the students' work. Watson referred to these adapted elements as the eight elements of a beginning inference framework. These eight elements of beginning inference were discussed with the teacher early in the study to ensure her awareness of them as she taught a series of four lessons using the *TinkerPlots* software. This study revealed that the use of this framework along with the software provided "an appropriate and valuable resource for introducing beginning inference to middle school students” (p. 80). In particular, the six elements of Hypothesis Generation, Summary, Shift, Signal, Spread, and Individual Case were supported by the software.

In the initial phase of an ongoing four-year study of the teaching of statistical inquiry, Makar and Rubin (2009) worked with four primary school teachers. Analysis of this initial phase

provided valuable insight into teachers' use of informal inference in teaching statistical inquiry. Makar and Rubin described informal statistical inference as "a reasoned but informal process of creating or testing generalizations from data, that is, not necessarily through standard statistical procedures" (p. 85). Initial concepts that Makar and Rubin saw as essential for informal inferential reasoning were:

- notion of uncertainty and variability articulated through language that broke from the mathematical convention of claims of certainty;
- reliance on the concept of aggregate (as opposed to individual points) through the use of generalizations about the group;
- acknowledgement of a mechanism or tendency that extended beyond the data at hand; and
- evidence for reasoning based on purposeful use of data. (p. 85)

From these initial concepts, Makar and Rubin then developed three principles they believed were essential to informal statistical inference:

> (1) *generalization*, including predictions, parameter estimates, and conclusions, that extend *beyond describing the given data*; (2) the use of data *as evidence* for those generalizations; and (3) employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn. (p. 85)

The principle of generalization requires the ability to reason about the population based on the available data. These generalizations "may be used to either generate hypotheses or evaluate them," (p. 86) without necessarily using statistical tests. The three essential principles of Makar and Rubin's framework were used to analyze lessons of statistical inquiry in primary classrooms.

This differs from the elements of reasoning in the framework used by Pfannkuch and Watson at the secondary level, however, a common theme of analyzing the data in aggregate to generalize beyond the data at hand exists.

The frameworks proposed and used by Pfannkuch, Watson, and Makar and Rubin facilitate the design of instruction and the analysis of students' informal inferential reasoning. An informal inferential reasoning framework proposed by Zieffler, Garfield, delMas, and Reading (2008) focused explicitly on task design. In light of research documenting the difficulties students have with formal inferential reasoning and recent attempts to describe informal inferential reasoning, Zieffler et al. (2008) developed this framework "for designing tasks that can be used to study students' reasoning about statistical inference" (p. 41). This framework consisted of three informal inferential reasoning components:

- making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures and methods (e.g., p-value, *t*-tests);
- drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts, such as distribution or average; informal knowledge about inference such as recognition that a sample may be surprising given a particular claim; use of statistical language), to the extent that this knowledge is available; and
- articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples. (p. 45)

There are similarities between the three components of this framework and the three principles proposed by Makar and Rubin (2009). The first component (or principle) of each involved making predictions about populations and the last component (or principle) of each included articulating and supporting those predictions.

Zieffler et al. (2008) believed this framework led to the design of tasks that would allow students to integrate all three of these components. Specifically, along with this framework they suggested three categories of tasks to conduct research on informal inferential reasoning:

- estimate and draw a graph of a population based on a sample;
- compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled; and
- judge which of two competing models or statements is more likely to be true. (p. 47)

The components of the reasoning framework and the categories of tasks proposed by Zieffler et al. (2008) together provided a melding of the frameworks used by Pfannkuch, Watson, and Makar and Rubin. Pfannkuch and Watson used tasks corresponding to the second category of tasks as they compared samples of data presented in boxplots. Makar and Rubin's essential principles of making predictions or drawing conclusions based on informed analysis of sample data and then describing these predictions or conclusions are reflected in Zieffler et al.'s informal inferential reasoning components. The Zieffler et al. framework provided additional guidance for designing tasks with which to study students' informal inferential reasoning.

Implications for This Study

The research conducted in statistics education has detailed student misconceptions as they learn formal statistical inference as well as the underlying concepts of descriptive statistics, probability, and the sampling distribution. Seeking to understand how these misconceptions might be dispelled, researchers have recently focused their efforts on students' informal inferential reasoning. With an understanding of how informal inferential reasoning develops in introductory statistics students, the statistics education community can inform instruction that supports students in their understanding of formal statistical inference.

To reiterate, the questions this study was designed to answer were:

1) For students enrolled in an introductory statistics class:

a) Does their informal inferential reasoning develop?

b) If their informal inferential reasoning develops, what are the characteristics of this informal inferential reasoning as it develops?

2) What is the relationship between students' informal inferential reasoning and their formal inferential reasoning?

This study was designed to add to current statistical education research seeking to inform instruction that provides students with a deeper understanding of statistical inference. The research area of informal statistical inference and the development of students' informal inferential reasoning holds potential for assisting students in deepening their understanding of what they can infer from data throughout their study of introductory statistics. With this understanding, the formalities and interpretation of formal statistical inference may no longer be a chasm but rather the next step.

Chapter 3 – Methods

Research in statistics education indicates that traditional instruction does not adequately support the development of the statistical reasoning students need for formal statistical inference. This study provided students with opportunities to engage in informal statistical inference as they progressed through their introductory statistics class. The focus of this study involved investigating the development of students' informal inferential reasoning as well as the nature and strength of the relationship between their informal and formal inferential reasoning.

Research Design

This was a mixed methods study conducted with students enrolled in introductory statistics classes taught in secondary schools. With the focus on understanding the development of students' informal inferential reasoning, the study began with all students in the statistics classes taking a pretest assessment containing informal statistical inference questions. This indicated the level of students' informal inferential reasoning at the beginning of their study of introductory statistics. All students were provided with the opportunity to engage in informal inferential reasoning as they took part in three informal statistical inference classroom activities over the course of the school year. These classroom activities also provided a starting point for discussion in the task-based interviews (Goldin, 2000) that followed them. A series of four task-based interviews with seven pairs of students were conducted to examine the development of the interviewees' informal inferential reasoning. The first three task-based interviews included informal statistical inference tasks with each one following the related classroom activity on the same topic. The fourth task-based interview contained formal statistical inference tasks and took place near the end of the school year. The study culminated with all students taking a posttest assessment of both informal and formal statistical inference questions.

Activities and task-based interviews.

The classroom activities consisted of three informal statistical inference activities which were observed by the researcher as they were carried out by the classroom teachers. The three activities took place in the following order in conjunction with the progression of the class curriculum: (1) comparing distributions of data; (2) a sampling and probability exploration; and (3) a sampling distribution activity (Appendix A). These corresponded to the three tasks of the modified task design framework of Zieffler et al. used for this study. These activities were designed to engage students in informal inferential reasoning as they learned introductory statistics.

Following each of these three classroom activities, task-based interviews were conducted with seven pairs of students. The pairs of students remained constant throughout the interviews. These interviews began with a recall of the classroom activity to gain insight into what these students learned from the activity. They were then asked to complete other problems in that same topic area to probe how their informal inferential reasoning was developing as they compared distributions of data, explored sampling and probability, and worked with the sampling distribution. The key statistical concepts for each task-based interview and the components of informal inferential reasoning are described in the data analysis section. A fourth task-based interview focused on formal statistical inference.

The structure of the task-based interviews followed the principles and techniques proposed by Goldin (2000). These included task-based interviews that (1) were designed specifically to answer the research questions, (2) included tasks with appropriate content for students' to grasp, (3) were structured based on key statistical concepts that gave students a variety of ways to demonstrate their understanding, (4) included explicit interview protocol

which allowed students to think about their responses without critiquing the correctness of their responses, and (5) involved students in free problem solving while they interacted with another student. Based on his own work as well as the work of others, Goldin believes that these principles and techniques "provide a solid foundation for optimizing systematically the research information gathered in interviews" (p.540). The classroom activities and interview tasks were based on topic areas as they were introduced to the students in their classes and designed with multiple parts that increased in complexity.

The first classroom activity and interview task used in this study involving comparing distributions of data was based on the theory of Ben-Zvi (2004) that comparing distributions will help students progress from a local perspective, within a data set, to a global perspective of describing variability between data sets. Watson and Moritz (1999) found that students who were able to see several aspects of data sets working together as a whole were best poised to make inferences when comparing those data sets. Their investigations included data sets of the same size and data sets of different size requiring a proportional understanding of variation and distribution. Comparing distributions of data provides students with the opportunity to gain a deeper understanding of the basic concepts of descriptive statistics that may foster the development of their informal inferential reasoning.

Inclusion of the sampling and probability activity and task was supported by the work of Konold et al. (2011). These researchers theorized that giving students the opportunity to estimate the probability of an event that cannot be summarized with a theoretical probability (e.g., unlike the probability of obtaining a sum of seven when tossing two die) supports their informal inferential reasoning by providing a conceptual understanding of the uncertainty that exists and a level of confidence in their inferences. Estimating in this manner also provides students with the

opportunity to consider the importance of random samples and large samples. Konold et al. found students lacked an understanding of these concepts in their investigation.

The sampling distribution activity and task was supported by the work of Saldanha and Thompson (2002) who found that even after instruction on the sampling distribution, students tended to compare the results from a sample to the distribution of the original population rather than to the sampling distribution. This activity and task are designed to support what they call a "multiplicative" conception of sample where the distinction between the population distribution, the distribution of a single sample taken from the population, and the distribution of the sample statistics of many samples is understood. Making an informal inference by examining where a sample statistic is situated in comparison to all such samples in the sampling distribution may help in developing this multiplicative conception of sample. This multiplicative conception of sample has the potential to support students' informal inferential reasoning and is necessary for formal statistical inference.

First classroom activity and interview task: Comparing distributions of data.

The first classroom activity and task-based interview involved students in informal inferential reasoning by asking them to make an inference about two populations based on sample data. This activity and interview followed the students' study of descriptive statistics including graphical displays of data (e.g. histograms and boxplots), measures of center (mean and median), and measures of variability (range and standard deviation).

In the first classroom activity, students compared distributions of data displayed in boxplots. These increased in complexity as students first considered measures of center and then, in addition, the variability in the data. These activity tasks were modified from comparing distribution questions developed by Garfield, Zieffler and Lane-Getaz (2005) as part of the

Adapting and Implementing Innovative Material in Statistics (AIMS) Project website (Garfield, delMas, & Zieffler, 2007). The website contains student activities aligned with the Guidelines for Assessment and Instruction in Statistics Education (GAISE) for teaching introductory statistics courses. This first classroom activity concluded with a metacognitive question about the important aspects they took into consideration while completing the activity. The task-based interview (Appendix B) following this classroom activity had students comparing distributions of data displayed in histograms. There were five parts to this task, each comparing data of children's test scores from two classes. These comparisons were modified from questions used by Watson and Moritz (1999). The student pairs were asked if the classes scored equally well or if one of the classes scored better (see Appendix B for Interview Protocol). The first part required a comparison of the measure of center with one of the classes clearly scoring better. The second part also required a comparison of measures of center, however, the students had to take the shape (one was skewed right and one was skewed left) into consideration. The third part added complexity as students were shown two distributions with the same mean but different variability. With yet another layer of complexity, the fourth part showed two classes of different size in which students had to reason proportionally in determining which class performed better. To complete the fifth and final comparison, students needed to combine their proportional reasoning with the concept of variability.

Second classroom activity and interview task: Sampling and probability.

The second classroom activity and task-based interview had students inferring about an unknown probability by collecting their own data. This followed their study of random sampling (sampling methods generating samples representative of the population that avoid bias) and, the Law of Large Numbers (as the number of independent repeated trials increases, the relative

frequency approaches the probability of the event), and basic probability rules (e.g. the probabilities in a model sum to one and the probabilities of complements).

During the classroom activity, students were working with identically-shaped, small plastic pigs which were tossed to determine the probability that the pigs would land on their backs. Students likely entered into this activity with intuitions about the chances of the pigs landing on their feet, their sides, or on their backs. However, the probabilities of these were not obvious and could not be clearly modeled with theoretical probabilities. Students estimated the probability that a pig would land on its back with data they collected. This classroom activity and the interview tasks were modeled after the Bone Problem used in the Konold et al. study (2011). In the first part of the task-based interview (Appendix C), students were asked to estimate the probability that a Monopoly house would land upright when it was tossed (see Appendix C for Interview Protocol). The students again had the opportunity to collect data. Once they completed their data collection and estimated this probability, I revealed results I collected from 1,000 tosses of the Monopoly houses. This was done to determine the extent of their understanding of the Law of Large Numbers. For the second part of the interview, students were shown a container of multi-colored beads and asked to estimate the number of green beads. For sampling, the students had a slotted paddle that drew samples of 32 beads at a time. This added complexity as students had to think about what method would provide random samples and then determine how these samples would provide information about the number of green beads in the container. I asked students to represent their estimate as a proportion or percentage of green beads in addition to the number of green beads. This allowed me to return to this proportion or percentage of beads in the fourth interview when students once again worked with the beads.

Third classroom activity and interview task: The sampling distribution.

The third classroom activity and task-based interview involved students working with sampling distributions. They made informal inferences about a population based on how samples of data of a particular size compared to all samples of data of that size. This followed students' study of sampling distributions during which they had been exposed to the normality of sampling distributions and to the effects of sample size on the variability of a sampling distribution (the larger the sample size, the less the variability in the sampling distribution).

In the classroom activity on sampling distributions, students first predicted what the sampling distribution would look like by choosing from several graphs of distributions and answered questions about the effect of sample size on the variability. This was an activity developed and used by Chance, delMas, and Garfield (2004). Students then were asked to make inferences by situating a sample statistic in relation to the sampling distribution. This activity served as an informal method to test a hypothesis. The task-based interviews began with a review of the Chance, et al. classroom activity (Appendix D). Students were then asked to talk about how they chose the graphs and what role variability played in their choices (see Appendix D for Interview Protocol). In the second part of the task-based interview, students viewed a simulation for building sampling distributions using the *Fathom* software. The students were then presented with three sampling distributions generated from the same Random Rectangles simulation activity in *Fathom*. The sample size increased from five to ten and then to 25 as the distributions of average areas were graphed. Students were asked which average areas would be likely and which would be rare or unlikely based on each of the sampling distributions. The interviews concluded by returning to the Tossing Monopoly Houses Activity from the second task-based interview to test the hypothesis that a Monopoly hotel had the same probability of landing

upright as a house. Students were shown a sampling distribution of sample proportions of houses landing upright generated from 200 samples of 10 houses to assist them in making this informal inference.

Fourth interview task.

The final task-based interview (Appendix E) encompassed formal statistical inference by asking students to interpret confidence interval estimates of population parameters and interpret the results of a hypothesis test (see Appendix E for Interview Protocol). This followed classroom instruction on these topics.

In this interview, students returned to the container of beads by first viewing 10 confidence intervals of the proportion of red beads constructed from random samples. They were asked what these confidence intervals revealed about the proportion of red beads in the container. They then took a sample to construct their own confidence interval and were asked to interpret the interval. The interview concluded with students conducting a hypothesis test using their sample of red beads to determine if they agreed or disagreed with a conjecture made about the proportion of red beads in the jar.

Pre/posttest assessment.

The study included a quantitative pre and posttest assessment of statistical inference questions to evaluate students' inferential reasoning (Appendix F). Ten of the items (questions 1, 2, 7, 8, 13, 14, 15, 18, 19, 21) came from or were modified from the Comprehensive Assessment of Outcomes in Statistics (CAOS 4) developed by delMas, Garfield, Ooms, and Chance (2007). The CAOS 4 multiple-choice assessment exceeded acceptable levels for internal consistency with a Cronbach's alpha coefficient of 0.82 and was used with a sample of 763 introductory statistics students enrolled in 20 different two-year and four-year colleges and universities from

14 different states in the United States. Seven additional items (questions 5, 9, 10, 16, 17, 20, 22) came from or were modified from the assessment item data base of the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project of which the CAOS 4 was a part. This project was funded by the National Science Foundation to develop "reliable, valid, practical, and accessible" (delMas, Ooms, Garfield, & Chance, 2006) items to assess students' statistical literacy, reasoning, and thinking. The data base was formulated over a two year period led by delMas, Ooms, Garfield, and Chance as the principal investigators with an advisory board of 12 experts in the field of statistics education. Together they revised and tested items improving validity. Four more of the pre/posttest items (questions 3, 4, 11, 12) were recommended for use in assessing students' informal inferential reasoning as part of the Zieffler et al. (2008) task framework which was used for this study. The remaining item (question 6), I designed based on the Spinner Problem used by Konold et al. (2011). In the Spinner Problem, students estimate an unknown probability of stopping on a certain color based on many spins of the needle.

The pretest was administered near the beginning of the school year and included informal statistical inference questions. This gave information about their informal inferential reasoning based on the statistics and data analysis they had experienced in mathematics classes prior to this introductory statistics class.

The study concluded with a posttest administered near the completion of the academic school year which contained both informal and formal statistical inference questions. This allowed me to determine if students' informal inferential reasoning had developed over the academic year and the nature of this development by examining the three types of informal inferential reasoning questions in this assessment. Corresponding to the three classroom

activities and the first three task-based interviews, these three types of questions were (1) comparing two samples of data (questions 1 – 4), (2) sampling to estimate a probability (questions 5 – 8), and (3) inferring about a population based on data from a single sample (questions 9 – 12). In addition, the posttest scores allowed me to determine if there was a correlation between students' informal and formal inferential reasoning and the nature of the correlation based on the three types of informal inferential reasoning questions.

Setting and Participants

The students taking part in this study were enrolled in introductory statistics courses for college credit in their high schools. These students were either in their 11^{th} or 12^{th} grade year and had successfully completed at least the first two courses of the three mathematics courses required for high school graduation. Therefore, each student had the required algebra background for their study of introductory statistics. These students came from one of eight statistics classes taught by four different high school mathematics teachers from two high schools, Deerfield High School and Rosemont High School (pseudonyms). There were four such classes at each high school with each teacher teaching two classes. These statistics classes met for approximately three and one-half hours each week for the 40-week school year beginning in September, 2011.

At Deerfield High School, the 64 students were taking their introductory statistics class through a local college. The introductory statistics course curriculum included descriptive statistics, bivariate relationships, randomness and study design, probability, sampling distributions, confidence intervals, and hypothesis tests, respectively. The DeVeaux, Velleman, and Bock (2009) textbook was used for this course. Students successfully completing the course earned four college credits. The 72 students at Rosemont High School were taking their statistics

class through a different local college. The curriculum covered descriptive statistics, randomness and study design, probability, sampling distributions, confidence intervals, hypothesis tests, regression, and analysis of variance, respectively. The Moore, McCabe, and Craig (2009) textbook was used for this course. Students successfully completing this course earned six college credits.

The teachers of these introductory statistics classes were experienced teachers with varying levels of experience teaching statistics. Among the teachers at Deerfield High School, one had been teaching for 13 years and teaching statistics for 12 of those years; the second teacher had been teaching for five years and statistics for four of those years. At Rosemont High School, one of the teachers had been teaching statistics for six of a total of 11 years. The second teacher had been teaching for five years and was teaching the statistics course at Rosemont for the first time. This teacher had previous experience teaching these statistics courses as a college teaching assistant.

I conducted a preliminary meeting with the teachers from each high school in June of 2011to give them a brief overview of the study and to secure their interest in participating. In early September of 2011, I met with the teachers once again to discuss the implementation of the pretest, the three classroom activities, and the posttest. I explained that I would administer both the pre and posttest. I did this to ensure that the testing conditions remained similar in all classrooms. The teachers agreed to administer the classroom activities while I observed. I was not focusing on the classroom teaching during these observations; and I wanted these activities to be a part of their regular instruction with their classroom teacher. We discussed the three informal statistical inference activities, the key statistical concepts related to each, and the timing of these activities to coordinate with the curriculum. I emphasized that my purpose was to gain

insight into how the students reasoned as they worked through the activities and participated in the classroom discussion.

A pair of students from each of the eight classes was asked to take part in the task-based interviews and all eight student pairs were selected with the assistance of the classroom teachers. Eight pairs were chosen to allow for possible attrition due to unforeseen circumstances preventing these students from taking part in the task-based interviews. I anticipated that a minimum of four of these pairs would complete the task-based interviews for the study and seven of the student pairs completed the study. The pairs of students represented a range of prior achievements in mathematics and each student in a pair had different prior achievement in mathematics. Students were categorized as high, middle, or low achieving and, for example, a high achieving student was paired with a middle-level achieving student. I elicited the teachers' assistance in selecting students with good attendance records, who are verbal, and would work well together. These criteria provided the greatest degree of assurance that students would have taken part in the classroom activities prior to the interviews and would express their understandings of the key statistical concepts and their informal inferential reasonings. This was beneficial not only in assessing the students' understandings and reasonings in each interview but also their progression over all of the interviews.

Data Collection

Each of the three informal statistical inference classroom activities was observed by the researcher in the eight classrooms. Brief notes on the implementation of the activities were recorded including students' insights, expected and unexpected, as they worked through the activities. Students' written work on these classroom activities was collected. Following each of the three classroom activities, the task-based interviews with the seven pairs of students were

videotaped. After each interview, a memo to file was written with initial impressions of students' work. A summary of what transpired along with any unusual or unforeseen student responses was recorded. The video tapes of students' responses were transcribed. There were a total of four interviews conducted with each pair of students. The pre and post quantitative assessments were administered by the researcher to ensure that conditions for each administration were consistent. The students completed the assessments individually with no assistance from their peers or teacher. Calculators were available to the students; however, the inferential reasoning nature of the questions were such that the calculators were of little use.

Data Analysis

Task-based interviews.

I first read the transcriptions of all task-based interviews, coding for the key statistical concepts and the three main principles of informal statistical inference described by Makar and Rubin (2009). In the Comparing Distributions task-based interviews, I analyzed students' responses to determine if they: (1) used means to compare distributions; (2) used variation to compare distributions; (3) recognized the effects of skewness on the mean; and (4) reasoned proportionally when distributions were of different size. In the Sampling and Estimating Probability task-based interviews, I analyzed students' responses to determine if they: (1) took random samples; and (2) used the Law of Large Numbers. In the Sampling Distribution task-based interviews, I analyzed students' responses to determine if they: (1) distinguished between the sampling distribution (approximately normal distribution of statistic) and the population distribution; and (2) recognized that the variability of the sampling distribution was less than that of the population and decreased as the sample size increased.

The overall purpose was to understand how these students' informal inferential reasoning was developing as they worked through the assigned tasks. Students' responses as they recalled the classroom activities and while completing the three informal statistical inference tasks were analyzed to determine if they were able to make informal inferences that included the three main principles of informal statistical inference described by Makar and Rubin (2009): (1) *generalization*, including predictions, parameter estimates, and conclusions, that extended *beyond describing the given data*; (2) the use of data *as evidence* for those generalizations; and (3) employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn (p. 85).

Data from the final task-based interviews were analyzed to determine if students were able to properly interpret the results of formal statistical inference problems involving confidence intervals and hypothesis tests. In addition, the transcriptions were coded for evidence of the use of the key statistical concepts from each of the three previous interview tasks.

In the second phase of analysis, I examined each part of the task-based interviews, one at a time, for all seven pairs of interviewees so I could distinguish the similar/different conclusions drawn by the interviewees and the similar/different reasonings they gave for their conclusions. In the third phase of analysis, I grouped the parts of each task by these similar/different conclusions to add or combine coding for the reasonings the students expressed or exhibited. In a final phase of analysis, I coded the video recordings by parts of each task-based interview which allowed me to repeatedly review these video recordings for confirming and disconfirming evidence as I wrote the final analysis.

Analysis of the first three informal inferential reasoning task-based interviews allowed me to answer both components of the first research question about the development of students'

informal inferential reasoning. These questions posed whether this reasoning developed and, if so, what were its characteristics. Analysis of the fourth formal inferential reasoning task-based interview helped in answering the second research question about the relationship between students' informal and formal inferential reasoning.

Classroom activities.

Data from the classroom activities served to confirm that students had the opportunity to take part in informal statistical inference activities. Differences in students' statistical reasoning based on the high school attended become evident in the analysis of the quantitative assessments and task-based interviews. A review of students' written work during the classroom activities helped in understanding these differences.

Quantitative assessment.

The subscores outlined in Table 1 were used in the analysis of the pre and posttest. All scores were based on the total number of correct responses to the assessment questions. The pretest only included informal statistical inference questions and the overall score on the posttest was divided into an informal statistical inference (ISI) subscore and a formal statistical inference (FSI) subscore measuring students' inferential reasoning. Students' ISI score, for both the pre and posttests, was further divided into COMP, PROB, and SAMP subscores measuring students' informal inferential reasoning when comparing two samples of data, when sampling and estimating a probability, and when inferring about a population based on a sample of data, respectively.

Table 1

Subscores of Assessment Questions

Subscore		Questions
	Measures of Informal Inferential Reasoning – Pre and Posttest	
ISI	overall	1 - 12
COMP	when comparing two distributions of data	1 – 4
PROB	when sampling and estimating a probability	5 – 8
SAMP	when inferring about a population based on a sample of data	9 - 12
FSI	Measure of Formal Inferential Reasoning – Posttest Only	13 - 22

Analysis of initial informal inferential reasoning.

The pretest scores gave an indication of students' informal inferential reasoning at the beginning of their introductory statistics class. Nested analyses of variance for the pretest ISI scores and pretest subscores, COMP, PROB, and SAMP, were conducted to determine if differences existed due to high school attended, classroom teacher, or individual class. It was determined that differences existed in the overall informal statistical inference (ISI) scores and in the subcategories of COMP and PROB by high school. Therefore, these analyses of pretest scores were also conducted by high school. The pretest scores of the 14 interviewees were analyzed and compared to the remaining students to determine if their initial informal inferential reasoning differed from the other 122 introductory statistics students. This was important for determining if these 14 interviewees were representative of the 136 introductory statistics students. I anticipated there would be no differences.

Analysis of posttest informal and formal inferential reasoning.

The posttest scores gave an indication of students' informal and formal inferential reasoning near the completion of their introductory statistics class. Nested analyses of variance

for the posttest ISI scores and posttest subscores, COMP, PROB, and SAMP, were conducted to determine if differences existed due to high school attended, classroom teacher, or individual class. It was determined that differences existed in the informal statistical inference (ISI) scores, in the subcategory of PROB, and in the overall statistical inference (SI) score by high school. Therefore, these analyses of the posttest scores were also conducted by high school. The posttest scores of the 14 interviewees were analyzed and compared to the remaining students to determine if their inferential reasoning differed from the other 122 introductory statistics students. This was done to determine if the task-based interviews had an impact on these students' inferential reasoning. I anticipated that the interviewees would score better on the posttest due to their work in the task-based interviews.

Analysis of change in informal inferential reasoning.

To determine whether students' informal inferential reasoning had improved due to their regular classroom instruction in introductory statistics, matched pairs *t*-tests or Wilcoxon Signed-rank tests (depending on normality) were conducted on all students' informal statistical inference scores and subscores from the pre and posttests. This helped to answer the first component of the first research question as to whether students' informal inferential reasoning had developed. Analyses of the subscore categories offered insight into the characteristics of students' informal inferential reasoning. This helped to answer the second component of the first research question about the characteristics of the development of students' informal inferential reasoning.

Since differences were detected by high school for the pre and posttest scores and subscores, analyses of the change in informal inferential reasoning were also conducted by high school to determine if these differed by high school. To determine if the 14 interviewees

benefitted from the task-based interviews, their changes in ISI, COMP, PRB, and SAMP scores were compared to those of the other students.

Analysis of the relationship between informal and formal inferential reasoning.

To determine if students' informal inferential reasoning was related to their formal inferential reasoning, the correlation between ISI and FSI from the posttest results was analyzed for all students. Similar correlations were analyzed with COMP, PROB, and SAMP and FSI to determine the relationship between students' informal inferential reasoning in each of the categories of questions and their formal inferential reasoning. This assisted with answering the second research question about the relationship between students' informal and formal inferential reasoning.

Since differences were detected by high school for the pre and posttest scores and subscores, correlations between informal and formal inferential reasoning were also conducted by high school. For the 14 interviewees, the correlations between ISI and FSI from the posttest results and between COMP, PROB, and SAMP and FSI were also run and compared to the corresponding correlations for all other students.

Among all of the students participating in the study, two additional groups were of particular interest when analyzing the relationship between informal and formal inferential reasoning. The first was the group of students showing a substantial improvement in their informal inferential reasoning from the pretest to the posttest and the second group consisted of those students who were strong informal inferential reasoners on the pretest and the posttest. The same correlations between informal and formal inferential reasoning were calculated separately for each of these groups of students. This gave more insight into the relationship between informal and formal inferential reasoning by determining if, at the completion of the introductory

statistics course, strong informal inferential reasoning correlated to strong formal inferential reasoning. Chapters 4 and 5 detail the quantitative and the qualitative analyses, respectively.

Chapter 4 - Quantitative Analysis

To begin, I will report on the pre and posttest results for the 136 introductory statistics students. As I will detail, some of these informal and formal statistical inferences scores differed for the two high schools. Therefore, the second segment of this analysis includes the pre and posttest results by high school. Lastly, I examined the pre and posttest results for the group of 14 interviewees who took part in four task-based interviews between the administrations of the pretest and the posttest. Within each of these segments I will discuss: (1) the pretest scores which provided information on students' initial informal inferential reasoning; (2) the posttest scores which provided information on students' informal and formal inferential reasoning at the end of their introductory statistics course; (3) the change in students' informal inferential reasoning from the pretest to the posttest; and (4) the relationship between their informal and formal inferential reasoning with the analysis of the posttest scores.

The subscores outlined in Table 1 were used in the analysis of the pre and posttests. All scores were based on the total number of correct responses to the assessment questions. The overall score on the posttest was divided into an informal statistical inference (ISI) subscore and a formal statistical inference (FSI) subscore measuring both aspects of students' inferential reasoning.

Students' ISI scores, for both the pre and posttests, were further divided into COMP, PROB, and SAMP subscores. These subscores measured students' informal inferential reasoning when comparing two distributions of data, when sampling and estimating a probability, and when inferring about a population based on a sample of data, respectively.

Results

I first report on the results of the pretests with the overall informal statistical inference (ISI) scores and the scores in each subcategory (COMP, PROB, and SAMP) as an indication of students' initial informal inferential reasoning. I then report on the posttest results which include the formal statistical inference (FSI) scores.

Pretest results - students' initial informal inferential reasoning.

The pretest scores gave a measure of students' informal inferential reasoning at the beginning of their introductory statistics course. Table 2 displays the informal statistical inference scores on the pretest (Pre ISI) and its three subscores (Pre COMP, Pre PROB, and Pre SAMP).

Table 2

Pretest Scores for the 136 Students

	Pretest Scores (n = 136) *M (SD)*	Percentage Correct
Pre ISI	7.47 (1.92)	62.25%
Pre COMP	3.01 (0.89)	75.25%
Pre PROB	2.33 (1.10)	58.25%
Pre SAMP	2.13 (0.95)	53.25%

The mean score was 7.47 out of 12 with a standard deviation of 1.92 for the 136 student participants. This score was comprised of four questions about comparing distributions (M = 3.01, SD = 0.89), four questions about sampling and estimating a probability (M = 2.33, SD = 1.10), and four questions about inference about a population based on a sample of data (M = 2.13, SD = 0.95). Overall students achieved 62.25% correct on the pretest. This was comprised of 75.25% correct on the comparing distributions questions, 58.25% correct on the sampling and

estimating probability questions, and 53.25% on the inference based on a sample questions. It was not surprising that students scored highest on questions involving comparing two distributions of data. Students likely relied on their knowledge of measures of center, a topic included in their previous mathematics courses, in answering those questions. They would have had little, or possibly no experience, in sampling and estimating probabilities and inferring about a population based on a sample of data.

Posttest results – students' informal and formal inferential reasoning.

The posttest data were analyzed in three ways. I first summarized students' inferential reasoning on the posttests with their total statistical inference (SI) scores, the overall informal statistical inference (ISI) scores, the results in each subcategory of informal statistical inference (COMP, PROB, and SAMP), and the formal statistical inference (FSI) scores. This gave a measure of students' informal and formal inferential reasoning at the end of their introductory statistics course. Second, I examined the change in students' informal inferential reasoning as indicated by their responses to the 12 informal statistical inference questions on the posttest in comparison to their responses to the same 12 informal statistical inference questions on the pretest. Last, I analyzed the relationship between their posttest responses to the 12 informal statistical inference questions and the 10 formal statistical inference questions.

To culminate this portion of the analysis, I examined the relationship between informal and formal inferential reasoning for two subgroups of students. The first subgroup consisted of students above the 90^{th} percentile in their informal statistical inferences gain scores. This was explored to determine if students showing a substantial improvement in their informal inferential reasoning throughout their study of introductory statistics were strong formal inferential reasoners. The second subgroup consisted of students who scored above the mean in their

informal statistical inferences scores on both the pretest and the posttest. This was explored to determine if students who began their introductory statistics class as strong informal inferential reasoners and remained as such were strong formal inferential reasoners as well.

Posttest scores for the 136 students.

Students' responses to the 22 questions on the posttest provided a measure of their inferential reasoning at the end of their introductory statistics course. Table 3 displays the informal statistical inference scores, its three subscores, the formal statistical inference subscore, and the total statistical inference score. The three subscores of informal statistical inference, Post COMP, Post PROB, and Post SAMP, were used to measure students' informal inferential reasoning when comparing two distributions of data, when sampling and estimating a probability, and when inferring about a population based on a sample of data, respectively.

Table 3

Posttest Scores for the 136 Students

	Posttest Scores (n = 136) *Mean (SD)*	Percentage Correct
Post ISI	8.48 (1.87)	70.67%
Post COMP	3.21 (0.78)	80.25%
Post PROB	2.90 (1.02)	72.50%
Post SAMP	2.36 (0.94)	59.00%
Post FSI	4.66 (1.64)	46.60%
Post SI	13.14 (2.71)	59.73%

The mean informal statistical inference (ISI) score was 8.48 out of 12 with a standard deviation of 1.87 for the 136 introductory statistics students. This score was comprised of four questions about comparing distributions (M = 3.21, SD = 0.78), four questions about sampling and estimating a probability (M = 2.90, SD = 1.02), and four questions about inference based on

a sample of data (M = 2.36, SD = 0.94). The mean formal statistical inference (FSI) score was 4.66 out of 10 with a standard deviation of 1.64; and the total statistical inference (SI) mean score was 13.14 out of 22 with 2.71 as the standard deviation.

Overall students achieved 59.73% correct on the posttest. This was comprised of 46.60% correct on the formal statistical inference questions and 70.67% correct on the informal statistical inference questions. The informal statistical inference percentage was further broken down into 80.25% correct on the comparing distributions questions, 72.50% correct on the sampling and estimating probability questions, and 59.00% correct on the inference based on a sample questions. This was an indication that at the end of their introductory statistics course, overall, students were not strong formal inferential reasoners. However, the improvement in their informal inferential reasoning was significant. I will now discuss these changes in more detail.

Change in informal inferential reasoning for the 136 students.

To determine whether students' informal inferential reasoning improved over the school year due to their regular classroom instruction in introductory statistics, students' responses to the 12 informal inferential reasoning questions on the posttest were compared to their responses to the same questions on the pretest. In addition, tests were conducted on each of the subscores, COMP, PROB, and SAMP, to determine if there were particular categories in which students showed improvement. These scores are displayed in Table 4.

Table 4

Change in Informal Statistical Inference Scores for the 136 Students

	Pretest (n = 136) *M (SD)*	Posttest (n = 136) *M (SD)*	Mean Difference (Percentage Increase)	Cohen's *d*
ISI	7.47 (1.92)	8.48 (1.87)	1.01 (13.52%)**	0.53
COMP	3.01 (0.89)	3.21 (0.78)	0.20 (6.64%)*	NA
PROB	2.33 (1.10)	2.90 (1.02)	0.57 (24.46%)**	NA
SAMP	2.13 (0.95)	2.36 (0.94)	0.23 (10.80%)*	0.24

**$p < .01$, *$p < .05$

There was an increase of 1.01 points constituting a 13.52% increase in the overall informal statistical inference (ISI) scores for all students. This was comprised of an increase of 0.20 points or 6.64% for the comparing distributions (COMP) subscores, an increase of 0.57 points or 24.46% for the sampling and estimating a probability (PROB) subscores, and an increase of 0.23 points or 10.80% for the inference based on a sample (SAMP) subscores.

Using the Shapiro-Wilk and Skewness-Kurtosis tests for normality, the Post Comp and Post PROB distributions did not pass. Since the distributions for the Post COMP subscores and the Post PROB subscores were not normally distributed, Wilcoxon Signed-rank tests, where *p*-values were computed by an asymptotic normal distribution, were used to determine if there were significant increases in these two subscores. Paired *t*-tests were used to determine if there was a significant increase in students' overall informal inferential reasoning, ISI, and in the subcategory of SAMP. These analyses revealed that there were significant improvements in the overall ISI scores ($t(135) = 6.22$, $p < .01$) and the COMP ($z = 2.46$, $p < .05$), PROB ($z = 5.49$, $p < .01$), and SAMP ($t(135) = 2.24$, $p < .05$) subscores from pretests to posttests. This indicated that, in general, students' informal inferential reasoning improved due to their regular classroom instruction. Further, the Cohen's effect size ($d = 0.53$) for the improvements in ISI scores

suggested a moderate practical significance and the Cohen's effect size (d = 0.24) for the improvements in SAMP subscores suggested a low practical significance.

These analyses of the changes in students' informal inferential reasoning assisted in answering the first research question about whether students' informal inferential reasoning developed during their study of introductory statistics; and, if so, what were the characteristics of their informal inferential reasoning as it developed. Based on the quantitative results from the pre and posttests, there was an indication that students' informal inferential reasoning did develop due to regular classroom instruction. There may have been other confounding effects on this development such as the textbooks used for the courses or the use of manipulatives and computer simulations in these classes. Since the classroom instruction was not the focus of this study, the exact reasons for these significant improvements were unknown. However, evidence of the development of students' informal inferential reasoning could be seen in the increase in students' overall informal statistical inference scores as well as increases in each of its subcategories. Analyses of the subscore categories offered insight into the characteristics of students' informal inferential reasoning. Students improved on questions about comparing two distributions of data, sampling and estimating a probability, and inferring about a population based on a sample of data. The greatest improvement was realized in the subcategory of sampling and estimating a probability.

Relationship between informal and formal inferential reasoning for the 136 students.

To determine if students' informal inferential reasoning was related to their formal inferential reasoning, the correlations between informal statistical inference (ISI) and formal statistical inference (FSI) scores from the posttests were analyzed. Both Pearson's Correlation and Spearman's Rank Correlation were used since, as stated previously, the distributions for Post

COMP and Post PROB did not pass tests for normality. Pearson's Correlation and Spearman's Rank Correlation were also used for each informal statistical inference subscore, COMP, PROB, and SAMP, and formal statistical inference scores to determine the strength of the relationship between students' informal inferential reasoning in each of the subcategories of questions and their formal inferential reasoning. The results displayed in Table 5 are the Spearman's Rank Correlations. The significant results of Pearson's Correlations were consistent with these results in that Post ISI and Post PROB were significantly correlated to Post FSI ($p < .05$).

Table 5

Spearman's Rank Correlations between Informal and Formal Inferential Reasoning for *the 136 Students*

	Post FSI ($n = 136$)
Post ISI	.21*
Post COMP	.06
Post PROB	.20*
Post SAMP	.16

*$p < .05$

The posttest results revealed significant positive correlations between students' formal statistical inference (FSI) scores and their overall informal statistical inference (ISI) scores ($p < .05$) as well as between students' formal statistical inference (FSI) scores and their sampling and estimating a probability (PROB) subscores ($p < .05$). It should be noted that with a large sample size, a relatively small value for the correlation coefficient can be significant. However, this suggested that there was a modest relationship between students' informal inferential reasoning and their formal inferential reasoning. Furthermore, the informal inferential reasoning students exhibited when answering questions about sampling and estimating a probability had the strongest relationship to their formal inferential reasoning. These relationships were explored further in the qualitative analysis reported in the next chapter.

S***tudent subgroups.***

Among all of the students participating in the study, two groups of students were of particular interest when analyzing the relationship between informal and formal inferential reasoning: the group of students showing a substantial improvement in their informal inferential reasoning from the pretest to the posttest and the group of students who were strong informal inferential reasoners on the pretest and the posttest. Correlations between the informal statistical inferences subscores and the formal statistical inference scores were computed separately for each of these groups of students. This helped to give insight into the relationship between informal and formal inferential reasoning by determining if, at the completion of the introductory statistics course, strong informal inferential reasoning or a substantial improvement in informal inferential reasoning correlated to strong formal inferential reasoning.

Relationship between informal and formal inferential reasoning for students with substantial improvement in informal inferential reasoning.

To determine what would constitute a substantial improvement in informal statistical inference scores, I analyzed the distribution of the change in ISI scores for all 136 students. The distribution passed tests for normality (*M* 1.01, *SD* 1.89). I chose 4 points as constituting the substantial increase since it represented the 90^{th} percentile for the change in ISI scores. The results are displayed in Table 6 for the 12 students achieving at least a four point increase in their informal statistical inference (ISI) scores from the pretest to the posttest.

Table 6

Spearman's Rank Correlations between Informal and Formal Inferential Reasoning for Students with Substantial Improvement in Informal Inferential Reasoning

	Post FSI (n = 12)
Post ISI	-.35
Post COMP	-.62*
Post PROB	.02
Post SAMP	.03

*$p < .05$

There was a significant correlation ($p < .05$) between Post FSI and Post COMP for the group of students showing a significant improvement in their informal inferential reasoning from the pretest to the posttest. The Pearson's Correlations also resulted in Post COMP significantly correlated to Post FSI ($p < .05$). However, this correlation was negative. Upon further investigation, I found that this was due to nine of these 12 students scoring five or less out of 10 on the formal statistical inference questions while also scoring three or four out of four on the comparing two distributions of data questions. Therefore, for this small group of students, they scored relatively high on the comparing distributions questions while scoring relatively low on the formal statistical inference questions. This attributed to the negative correlation that existed between their Post COMP and Post FSI scores. In addition, 10 of these 12 students showing a significant improvement in their informal statistical inference scores attended Deerfield High School. The remaining informal statistical inference scores of Post ISI, Post PROB, and Post SAMP revealed no significant correlations to students' formal statistical inference scores. These results were consistent with the results for the 64 students at Deerfield, as will be reported in the posttest results by high school in the next segment.

Therefore, for students demonstrating substantial improvement in their informal inferential reasoning by the end of their introductory statistics class, there were no positive

correlations to their formal inferential reasoning as was expected. Instead, for this group of 12 students, those who were strong informal inferential reasoners when responding to comparing distribution questions were not strong formal inferential reasoners. This would indicate that a substantial improvement in informal inferential reasoning alone does not signify a strong correlation between informal and formal inferential reasoning.

Relationship between informal and formal inferential reasoning for strong informal inferential reasoners.

This group of students consisted of those scoring above the mean on the informal statistical inference questions on both the pretest and the posttest. The mean informal statistical inference scores were 7.47 on the pretest and 8.48 on the posttest. There were 46 students scoring at least eight out of 12 on the pretest and at least nine out of 12 on the posttest on these informal statistical inference questions. To provide additional information about these 46 students, their responses to the 12 informal inferential reasoning questions on the pretest and the posttest were compared to determine if there were significant increases in their informal statistical inference scores. These scores are displayed in Table 7.

Table 7

Change in Informal Statistical Inference Scores from Pretest and Posttest for Strong Informal Inferential Reasoners

	Pretest (n = 46) *M (SD)*	Posttest (n = 46) *M (SD)*	Mean Difference (Percentage Increase)	Cohen's *d*
ISI	9.20 (1.05)	10.17 (0.95)	0.97 (10.54%)**	NA
COMP	3.52 (0.59)	3.67 (0.52)	0.15 (4.26%)	NA
PROB	3.20 (0.78)	3.57 (0.62)	0.37 (11.56%)**	NA
SAMP	2.48 (0.86)	2.93 (0.83)	0.45 (18.15%)**	0.53

**$p < .01$

There was an increase of 0.97 points constituting a 10.54% increase in the overall informal statistical inference (ISI) scores for these 46 students. This was comprised of an increase of 0.15 points or 4.26% for the comparing two distributions of data (COMP) subscores, an increase of 0.37 points or 11.56% for the sampling and estimating a probability (PROB) subscores, and an increase of 0.45 points or 18.15% for the inferring about a population based on a sample of data (SAMP) subscores.

Using the Shapiro-Wilk and Skewness-Kurtosis tests for normality, the Pre SAMP and Post SAMP distributions were the only distributions to pass. Since the remaining distributions were not normally distributed, Wilcoxon Signed-rank tests, where *p*-values were computed by an asymptotic normal distribution, were used to determine if there were significant increases in the ISI, COMP, and PROB subscores. A paired *t*-test was used to determine if there was a significant increase in the subcategory of SAMP for these 46 students. These analyses revealed that there was a significant improvement in the overall ISI scores ($z = 4.90$, $p < .01$), the PROB subscores ($z = 2.65$, $p < .01$), and the SAMP subscores ($t(45) = 2.95$, $p < .01$, $d = 0.53$) from pretest to posttest. This indicated that, except for the subcategory of COMP, informal inferential reasoning for these 46 students improved due to their regular classroom instruction. Therefore, not only were these 46 students above average informal inferential reasoners initially, they also showed a significant improvement in their overall informal inferential reasoning. Further, the Cohen's effect size ($d = 0.53$) for the improvements in SAMP subscores suggested a moderate practical significance.

The formal statistical inference scores for these 46 strong informal inferential reasoners were compared to the remaining 90 introductory statistics students to determine if there was a significant difference. The Post FSI distributions for both of these groups of students passed the

Shapiro-Wilk and Skewness-Kurtosis tests for normality. Therefore, a non-pooled two-sample *t*-test was used to determine if there was a significant difference in Post FSI scores. The 46 strong informal inferential reasoners scored significantly higher on the formal statistical inference questions than the remaining 90 students, $t(80.1736) = 2.09$, $p < .05$, with the Cohen's effect size ($d = 0.39$) suggesting a low to mederate practical significance.

There was a significant correlation between these students' overall informal statistical inferences scores and their formal statistical inference scores, as shown in Table 8. This was largely due to the significant correlation between their inference based on a sample (Post SAMP) subscores and their formal statistical inference scores. The Pearson's Correlations also showed Post ISI and Post SAMP significantly correlated to Post FSI ($p < .05$). This was an indication that students who began as and remained strong informal inferential reasoners, particularly those strong in inferring about a population based on a sample of data, were also strong formal inferential reasoners.

Table 8

Spearman's Rank Correlations between Informal and Formal Inferential Reasoning for Strong Informal Inferential Reasoners

	Post FSI ($n = 46$)
Post ISI	.36*
Post COMP	-.14
Post PROB	.19
Post SAMP	.40**

**$p < .01$, *$p < .05$

Summarizing the results for these two additional subgroups, the 12 students demonstrating a substantial improvement in their informal inferential reasoning from the pretest to the posttest showed no positive correlations between their informal and formal inferential reasoning on the posttest. For the 46 students who were strong informal inferential reasoners,

there was a positive correlation between their informal and formal inferential reasoning. These 46 students also showed a significant improvement in their overall informal inferential reasoning. This may indicate that for correlations to exist between students' informal and formal inferential reasoning at the completion of their introductory statistics course, improvements in informal inferential reasoning may not be as important as initial informal inferential reasoning at the beginning of the introductory statistics course.

To summarize results for the 136 students who took the pretest and the posttest, the analyses showed that their informal inferential reasoning improved due to regular classroom instruction. Students improved on questions in all three of the informal statistical inference subcategories: comparing two distributions of data; sampling and estimating a probability; and inferring about a population based on a sample of data. The greatest improvement was realized in the subcategory of sampling and estimating a probability. Regarding the relationship between informal and formal inferential reasoning for the 136 students, the results suggested that there was a modest relationship between the two as evidenced by their responses on the posttests. The informal inferential reasoning students exhibited when answering questions about sampling and estimating a probability had the strongest relationship to their formal inferential reasoning.

The analyses reported thus far were conducted for the 136 introductory statistics students who took the pretest and the posttest. Differences were found in the pretest and posttest scores by high school, therefore I will now report on those results.

Results by High School

In this analysis of results by high school, I first report on the results of the pretests with the overall informal statistical inference (ISI) scores and the scores in each subcategory (COMP,

PROB, and SAMP). I then report on the posttest results by high school which included the formal statistical inference (FSI) scores.

Pretest results by high school - students' initial informal inferential reasoning.

The pretest scores provided a measure of students' informal inferential reasoning at the beginning of their introductory statistics course. These pretest scores were analyzed to determine if there were differences in students' performance by high school attended, classroom teacher, or individual classes taught by those teachers. The nested ANOVA analyses displayed in Table 9 revealed that the Pre ISI, Pre COMP, and Pre PROB pretest scores were significantly different at the high school level. There were no differences in pretest scores between students of the four teachers or of the eight individual classes taught by those teachers.

Table 9

Nested ANOVA Results for Pretest Scores

Source	*df*	*MSE*	*F*	*p*
		Pre ISI		
School	1	42.677	12.580**	.001
Teacher	2	4.914	1.450	.239
Class	4	1.825	0.540	.708
		Pre COMP		
School	1	4.069	5.580*	.020
Teacher	2	0.684	0.940	.394
Class	4	1.766	2.420	.052
		Pre PROB		
School	1	9.009	7.880**	.006
Teacher	2	1.851	1.620	.202
Class	4	0.368	0.320	.863
		Pre SAMP		
School	1	2.293	2.490	.117
Teacher	2	0.023	0.030	.975
Class	4	0.356	0.390	.818

**$p < .01$, *$p < .05$

There was a difference by high school in the overall informal statistical inference scores, Pre ISI, ($p < .01$). This was accounted for by differences between the high schools in the

subcategory scores of Pre COMP ($p < .05$) and Pre PROB ($p < .01$). There was no difference detected in their Pre SAMP subscores.

Table 10 displays the informal statistical inference scores on the pretest (Pre ISI) and the three subscores (Pre COMP, Pre PROB, and Pre SAMP) for each of the high schools.

Table 10

Pretest Scores by High School

	Deerfield High School $n = 64$ *M (SD)*	Rosemont High School $n = 72$ *M (SD)*	Differences between High Schools
Pre ISI	6.83 (1.84)	8.04 (1.83)	1.21**
Pre COMP	2.80 (0.98)	3.19 (0.76)	0.39*
Pre PROB	2.03 (1.10)	2.60 (1.03)	0.57**
Pre SAMP	2.00 (0.85)	2.25 (1.02)	0.25

**$p < .01$, *$p < .05$

The 64 Deerfield students had a mean score of 6.83 out of 12 compared to 8.04 for the 72 Rosemont students in their overall informal statistical inference (Pre ISI) scores. For the informal statistical inference subcategories, Deerfield students had a mean score of 2.80 points out of 4 compared to 3.19 points for the Rosemont students for the comparing distributions (Pre COMP) subscores; a mean of 2.03 out of 4 compared to 2.60 for the Rosemont students for the sampling and estimating a probability (Pre PROB) subscores; and a mean of 2.00 out of 4 compared to 2.25 for the Rosemont students for inference based on a sample (Pre SAMP) subscores.

These results indicated that students at Rosemont High School were stronger informal inferential reasoners overall as they began their introductory statistics course. In particular, they were stronger informal inferential reasoners when answering questions that involved comparing

two distributions of data (Pre COMP) and sampling and estimating a probability (Pre PROB). This was likely due to instruction they received in their previous mathematics courses. No differences were found between the high schools on questions that involved inferring about a population based on a sample of data (Pre SAMP). As was stated in the prior analyses for the 136 students, they would likely have had limited experience with questions of this kind in their previous high school mathematics courses.

Posttest results by high school - students' informal and formal inferential reasoning.

The posttest scores were then analyzed to determine if there were differences in students' performance by high school attended, classroom teacher, or individual classes taught by those teachers. The nested ANOVA analyses displayed in Table 11 revealed that the Post ISI, Post PROB, and Post SI scores were significantly different at the high school level. Consistent with the pretest scores, there were no differences in posttest scores between students of the four teachers or of the eight individual classes taught by those teachers.

Table 11

Nested ANOVA Results for Posttest Scores

Source	*df*	*MSE*	*F*	*p*
		Post ISI		
School	1	16.971	4.850*	.029
Teacher	2	1.789	0.510	.601
Class	4	0.424	0.120	.975
		Post COMP		
School	1	1.903	3.120	.079
Teacher	2	0.633	1.040	.357
Class	4	0.273	0.450	.774
		Post PROB		
School	1	7.339	7.450**	.007
Teacher	2	1.891	1.920	.151
Class	4	0.245	0.250	.910
		Post SAMP		
School	1	0.001	0.000	.974
Teacher	2	1.990	2.230	.112
Class	4	0.117	0.130	.971
		Post FSI		
School	1	1.603	0.590	.443
Teacher	2	4.362	1.610	.204
Class	4	2.737	1.010	.404
		Post SI		
School	1	29.005	3.970*	.048
Teacher	2	11.496	1.580	.211
Class	4	3.048	0.420	.796

**$p < .01$, *$p < .05$

The nested ANOVA analysis revealed that differences existed by high school in the overall informal statistical inference scores, Post ISI, ($p < .05$), in the subcategory score of Post PROB ($p < .01$), and the overall statistical inference scores, Post SI ($p < .05$). Additionally, no differences were detected between high schools in formal statistical inference scores, Post FSI, or in the informal statistical inference subscores of Post COMP and Post SAMP.

Table 12 displays the informal statistical inference scores and subscores, the formal statistical inference scores, and the overall statistical inference scores for each of the high schools.

Table 12

Posttest Scores by High School

	Deerfield High School *n* = 64 *M (SD)*	Rosemont High School *n* = 72 *M (SD)*	Differences between High Schools
Post ISI	8.08 (1.95)	8.83 (1.74)	0.75*
Post COMP	3.08 (0.84)	3.33 (0.71)	0.25
Post PROB	2.63 (1.08)	3.15 (0.90)	0.52**
Post SAMP	2.38 (0.98)	2.35 (0.91)	0.03
Post FSI	4.58 (1.50)	4.74 (1.77)	0.16
Post SI	12.66 (2.50)	13.57 (2.84)	0.91*

*$p < .01$, *$p < .05$

The 64 Deerfield students had a mean score of 8.08 out of 12 compared to 8.83 for the 72 Rosemont students in their overall informal statistical inference (Post ISI) scores. For the informal statistical inference subcategories, Deerfield students had a mean score of 3.08 points out of 4 compared to 3.33 points for the Rosemont students for the comparing distributions (Post COMP) subscores; a mean of 2.63 out of 4 compared to 3.15 for the Rosemont students for the sampling and estimating a probability (Post PROB) subscores; and a mean of 2.38 out of 4 compared to 2.35 for the Rosemont students for the inference based on a sample (Post SAMP) subscores. The Deerfield students scored a mean of 4.58 points out of 10 compared to 4.74 points for the Rosemont students on the formal statistical inference (Post FSI) subscore. These combined to give the Deerfield students an overall statistical inference (Post SI) mean score of 12.66 points out of 22 compared to 13.57 for the Rosemont students.

These results indicated that at the end of their introductory statistics course, students at Rosemont High School were stronger inferential reasoners overall than students at Deerfield. This could be attributed to the fact that they were stronger informal inferential reasoners than

students at Deerfield High School. This, in turn, could be attributed to the Rosemont students demonstrating stronger informal inferential reasoning when responding to questions about sampling and estimating a probability.

The Rosemont students were also stronger informal inferential reasoners initially based on their pretest scores. Therefore, I tested the differences in informal statistical inference scores on the posttest by school with the pretest informal statistical inferences scores as a covariate. An analysis of covariance, $F(1,133) = 38.75$, $p = .000$, indicated that the informal statistical inference scores on the pretest had a significant effect on the informal statistical inference scores on the posttest. In addition, with the pretest informal statistical inference scores as a covariate, there was no significant difference between the posttest informal statistical inference scores by high school ($p = .55$). This was an indication that students' initial informal inferential reasoning was an important factor for their informal inferential reasoning at the end of the introductory statistics course.

There was no difference between schools in their formal inferential reasoning on the posttest. With an analysis of covariance, I tested the formal statistical inference scores by high school with the pretest and posttest informal statistical inference scores as covariates since differences were detected in both by high school. Still no difference was detected by high school in the formal statistical inference scores. Therefore, regardless of the differences that existed in informal inferential reasoning on the pretest and the posttest, there were no differences by high school in students' formal inferential reasoning at the end of their study of introductory statistics. With each high school averaging less than 50% on the formal statistical inference questions, this indicated that the mastery of formal statistical inference was not trivial and regular classroom instruction may not have provided the necessary support.

Change in informal inferential reasoning by high school.

For the 136 students in the study, their informal inferential reasoning improved overall and in each of the three subcategories when students' responses to the 12 informal inferential reasoning questions on the posttest were compared to their responses to the same questions on the pretest. Further analyses were conducted on these changes to determine if differences existed by high school. The changes in informal statistical inference scores for Deerfield students are displayed in Table 13 and for Rosemont students in Table 14.

Table 13

Change in Informal Statistical Inference Scores for Deerfield Students

	Pretest (n = 64) *M (SD)*	Posttest (n = 64) *M (SD)*	Mean Difference (Percentage Increase)	Cohen's *d*
ISI	6.83 (1.84)	8.08 (1.95)	1.25 (18.30%)**	0.66
COMP	2.80 (0.98)	3.08 (0.84)	0.28 (10.00%)	NA
PROB	2.03 (1.10)	2.63 (1.08)	0.60 (29.56%)**	0.55
SAMP	2.00 (0.85)	2.38 (0.98)	0.38 (19.00%)*	0.41

**p < .01, *p < .05

There was an increase of 1.25 points constituting an 18.30% increase in the overall informal statistical inference (ISI) scores for the Deerfield students. This was comprised of an increase of 0.28 points or 10.00% for the comparing distributions (COMP) subscores, an increase of 0.60 points or 29.56% for the sampling and estimating a probability (PROB) subscores, and an increase of 0.38 points or 19.00% for the inference based on a sample (SAMP) subscores.

Using the Shapiro-Wilk and Skewness-Kurtosis tests for normality, the Pre Comp and Post COMP distributions did not pass. Therefore, Wilcoxon Signed-rank tests, where *p*-values are computed by an asymptotic normal distribution, were used to determine if there was a significant increase in the COMP subscores. Paired *t*-tests were used to determine if there were

significant increases in Deerfield students' overall informal inferential reasoning, ISI, and in the subcategories of PROB and SAMP. These analyses revealed that there were significant improvements in the overall ISI scores ($t(63) = 5.19$, $p < .01$) and the PROB ($t(63) = 4.34$, $p < .01$) and SAMP ($t(63) = 2.32$, $p < .05$) subscores from pretests to posttests. This indicated that the Deerfield students' informal inferential reasoning improved due to their regular classroom instruction overall and in the subcategories of sampling and estimating a probability and inferring about a population based on a sample of data. Further, the Cohen's effect size ($d = 0.66$) for the improvements in ISI scores suggested a moderate to high practical significance, the Cohen's effect size ($d = 0.55$) for the improvements in PROB subscores suggested a moderate practical significance, and the Cohen's effect size ($d = 0.41$) for the improvements in SAMP subscores also suggested a moderate practical significance.

These results differed slightly for the Rosemont students.

Table 14

Change in Informal Statistical Inference Scores for Rosemont Students

	Pretest (n = 72) *M (SD)*	Posttest (n = 72) *M (SD)*	Mean Difference (Percentage Increase)	Cohen's *d*
ISI	8.04 (1.83)	8.83 (1.74)	0.79 (9.83%)**	0.44
COMP	3.19 (0.76)	3.33 (0.71)	0.14 (4.39%)	NA
PROB	2.60 (1.03)	3.15 (0.90)	0.55 (21.15%)**	NA
SAMP	2.25 (1.02)	2.35 (0.91)	0.10 (4.44%)	0.10

**$p < .01$, *$p < .05$

For the Rosemont student, there was an increase of 0.79 points constituting a 9.83% increase in the overall informal statistical inference (ISI) scores. This was comprised of an increase of 0.14 points or 4.39% for the comparing distributions (COMP) subscores, an increase

of 0.55 points or 21.15% for the sampling and estimating a probability (PROB) subscores, and an increase of 0.10 points or 4.44% for the inference based on a sample (SAMP) subscores.

Using the Shapiro-Wilk and Skewness-Kurtosis tests for normality, the Pre Comp, Post COMP, and Post PROB distributions did not pass. Therefore, Wilcoxon Signed-rank tests, where *p*-values were computed by an asymptotic normal distribution, were used to determine if there were significant increases in the COMP and PROB subscores. Paired *t*-tests were used to determine if there were significant increases in Deerfield students' overall informal inferential reasoning, ISI, and in the subcategory SAMP. These analyses revealed that there were significant improvements in the overall ISI scores ($t(71) = 3.64$, $p < .01$), with a Cohen's effect size ($d = 0.44$) suggesting a moderate practical significance, and the PROB ($z = 3.86$, $p < .01$) subscores from pretest to posttest. This indicated that the Rosemont students' informal inferential reasoning improved due to their regular classroom instruction overall and in the subcategory of sampling and estimating a probability.

Analyses of the pretest and posttest scores for the 136 introductory statistics students showed improvements in students' overall informal inferential reasoning as well as in each of the three subcategories. However, when broken down by high school, improvements in the overall informal inferential reasoning and in the subcategory of sampling and estimating a probability were the only significantly remaining for both high schools. Individually, students from neither high school showed an improvement when comparing two distributions of data and only Deerfield students showed an improvement when inferring about a population based on a sample of data.

The Pre COMP (comparing distributions of data) subscores were analyzed to determine if there may have been a ceiling effect preventing a significant improvement in that subcategory for

the high schools. Wilcoxon Signed-rank tests, where *p*-values were computed by an asymptotic normal distribution, revealed that on the pretest, Deerfield students scored significantly better on the COMP subscores than the PROB ($z = 3.85$, $p < .01$) and the SAMP ($z = 4.56$, $p < .01$) subscores. Similarly, the Rosemont students scored significantly better on the COMP subscores than the PROB ($z = 3.83$, $p < .01$) and the SAMP ($z = 5.55$, $p < .01$) subscores. This further analysis showed that both Deerfield and Rosemont students scored significantly better in this subcategory of comparing distributions than the other two subcategories on the pretest (refer to Table 10 for these subscores). This, indeed, appears to have created a ceiling effect since there were only four questions in the subcategory, not allowing for a significant improvement on the posttest for either high school.

Relationship between informal and formal inferential reasoning by high school.

To determine if students' informal inferential reasoning was related to their formal inferential reasoning, the correlations between informal statistical inference (ISI) and formal statistical inference (FSI) scores from the posttests were analyzed. These correlations were previously examined for the 136 students, then for each school since differences were found in the posttest scores of Post SI, Post ISI, and Post PROB by high school. This was done to determine if these differences in posttest scores by high school had an impact on the correlations between students' informal and formal inferential reasoning by high school. The Spearman's Rank Correlations for Deerfield High School and Rosemont High School are displayed in Table 15.

Table 15

Spearman's Rank Correlations between Informal and Formal Inferential Reasoning for Deerfield High School and Rosemont High School

	Post FSI	
	Deerfield (n = 64)	Rosemont (n = 72)
Post ISI	.06	.35**
Post COMP	-.03	.13
Post PROB	.06	.33**
Post SAMP	.03	.27*

**p < .01, *p < .05

The Pearson's Correlations also resulted in Post ISI and Post PROB significantly correlated to Post FSI (p < .01) for Rosemont students along with no significant correlations between informal statistical inference scores and formal statistical inference scores for Deerfield students. For Rosemont students, the same significant positive correlations existed as those for the 136 students. These results indicated that the significant positive relationship found for the 136 students between their informal inferential reasoning, specifically the reasoning they exhibited when answering questions about sampling and estimating a probability, and their formal inferential reasoning could be attributed to the students at Rosemont High School. In addition, for the Rosemont students, there was also a significant positive correlation between their formal statistical inference (FSI) scores and their inference based on a sample (SAMP) subscores (p < .05).

For the Deerfield students, their informal inferential reasoning did not correspond to their formal inferential reasoning. However, for the Rosemont students, the stronger informal inferential reasoners they were, particularly when reasoning about sampling and estimating a probability and when inferring about a population based on a sample of data, the stronger formal inferential reasoners they were as well. There were indications of what might be causing these

differences between the schools in the analysis of the task-based interviews reported in the next chapter. I next report on the pretest and posttest results for the 14 students who took part in those four task-based interviews throughout the school year.

Results for Interviewees

A pair of students from each of the eight classes were asked to take part in the task-based interviews and selected with the assistance of their classroom teachers. Of the initial eight pairs of students, seven pairs completed all four task-based interviews. I elicited the teachers' assistance in selecting students with good attendance records, who were verbal, and worked well together. In addition, I asked the teachers to recommend students for each pair that had differing prior achievement in mathematics. These criteria proved to be beneficial in assessing the students' understandings and reasonings in each interview and their progression over all of the interviews. In this analysis of results for the 14 interviewees, I will first report on the results of the pretests with the overall informal statistical inference (ISI) scores and the scores in each subcategory (COMP, PROB, and SAMP). I will then report on their posttest results which included the formal statistical inference (FSI) scores.

Pretest results – interviewees' initial informal inferential reasoning.

I anticipated that the pretest scores for the 14 interviewees would not differ from the other 122 students. For analysis purposes, their total pretest scores, Pre ISI, and the pretest subscores, COMP, PROB, and SAMP, were compared to those of the other students to determine if differences existed in their informal inferential reasoning at the beginning of their study of introductory statistics. The results are displayed in Table 16.

Table 16

Pretest Scores for Interviewees

	Interviewees $n = 14$ *M (SD)*	All Other Students $n = 122$ *M (SD)*	Differences between Interviewees and All Other Students	Cohen's *d*
Pre ISI	7.64 (2.06)	7.45 (1.92)	0.19	0.10
Pre COMP	2.79 (0.97)	3.03 (0.88)	0.24	0.26
Pre PROB	2.64 (1.15)	2.30 (1.09)	0.34	0.30
Pre SAMP	2.21 (0.97)	2.12 (0.95)	0.09	0.09

The 14 interviewees had a mean score of 7.64 out of 12 compared to 7.45 for all other students in their overall informal statistical inference (Pre ISI) scores. The interviewees had a mean score of 2.79 points out of 4 compared to 3.03 points for all other students for the comparing distributions (Pre COMP) subscores; a mean of 2.64 out of 4 compared to 2.30 for all other students for the sampling and estimating a probability (Pre PROB) subscores; and a mean of 2.21 out of 4 compared to 2.12 for all other students for the inference based on a sample (Pre SAMP) subscores.

Non-pooled two-sample *t*-tests were used to determine if there were differences in the Pre ISI, Pre PROB, and Pre SAMP, scores for the interviewees. A Wilcoxon Rank-sum test was used for the Pre COMP subscores since this distribution did not pass tests for normality. As anticipated, there were no significant differences in the pretest scores and subscores between the interviewees and the other students.

Posttest results – interviewees' informal and formal inferential reasoning.

I anticipated that students taking part in the task-based interviews would score significantly better than the other students on the posttest due to their extra work during the task-

based interviews. Their scores and those for the remaining 122 students are displayed in Table 17.

Table 17

Posttest Scores for Interviewees

	Interviewees $n = 14$ *M (SD)*	All Other Students $n = 122$ *M (SD)*	Differences between Interviewees and All Other Students	Cohen's *d*
Post ISI	8.79 (1.76)	8.44 (1.89)	0.35	0.19
Post COMP	3.29 (0.61)	3.20 (0.80)	0.09	NA
Post PROB	2.93 (0.83)	2.90 (1.04)	0.03	NA
Post SAMP	2.57 (1.09)	2.34 (0.92)	0.23	0.23
Post FSI	5.57 (1.22)	4.56 (1.66)	1.01**	0.69
Post SI	14.36 (2.41)	13.00 (2.72)	1.36*	0.53

**$p < .01$, *$p < .05$

The 14 interviewees had a mean score of 8.79 out of 12 compared to 8.44 for all other students in their overall informal statistical inference (Post ISI) scores. The interviewees had a mean score of 3.29 points out of 4 compared to 3.20 points for all other students for the comparing distributions (Post COMP) subscores; a mean of 2.93 out of 4 compared to 2.90 for all other students for the sampling and estimating a probability (Post PROB) subscores; and a mean of 2.57 out of 4 compared to 2.34 for all other students for the inference based on a sample (Post SAMP) subscores. The interviewees scored a mean of 5.57 points out of 10 compared to 4.56 points for all other students on the formal statistical inference (Post FSI) subscore. These combined to give the interviewees an overall statistical inference (Post SI) mean score of 14.36 out of 22 compared to 13.00 for all other students.

Their total posttest scores, Post SI, and all posttest subscores, ISI, COMP, PROB, SAMP, and FSI, were compared to those of the other students to determine if there was a significant difference due to their additional work with the interview tasks. As a reminder, these

interviewees worked as pairs throughout four task-based interviews. During the first interview, students compared distributions of data; during the second they sampled and estimated probabilities; during the third they inferred about populations based on samples of data; and in the fourth interview the student pairs worked on problems involving formal statistical inference. Non-pooled two-sample *t*-tests were used to determine if there was a difference in the Post SI, Post ISI, Post SAMP, and Post FSI scores for the 14 interviewees and the other 122 students. Wilcoxon Rank-sum tests were used for the Post COMP and Post PROB subscores since these distributions did not pass tests for normality. As noted in Table 15, these tests revealed that the overall statistical inference scores (Post SI), which included all informal and formal statistical inference questions, were significantly better for the interviewees ($t(17.05) = 1.79$, $p < .05$). Further, the Cohen's effect size ($d = 0.69$) suggested a moderate to high practical significance. This was largely due to the fact that the interviewees scored better on the formal statistical inference, Post FSI, portion of the posttest ($t(18.97) = 2.82$, $p < .01$). The Cohen's effect size ($d = 0.53$) for the interviewees' better FSI scores suggested a moderate practical significance. Since there were no differences in pretest scores and subscores between the 14 interviewees and the 122 other introductory statistics students, this may be an indication that their extra work during the task-based interviews benefitted their formal inferential reasoning at the culmination of the study. This may be due to the fact that these interviews took place following each of the classroom activities which gave the interviewees an additional opportunity to explore key statistical concepts. In addition, the interviews were designed according to the principles set forth by Goldin (2000) which included tasks with appropriate content for students to grasp, were structured based on key statistical concepts that gave students a variety of ways to demonstrate their understanding, included an explicit interview protocol that allowed students to think about

their responses without critiquing the correctness of their responses, and involved students in free problem solving while they interacted with another student. The interview tasks were also designed with multiple parts that increased in complexity. Together, the timing and the structure of the interviews may have supported the interviewees' formal inferential reasoning.

Change in informal inferential reasoning for interviewees.

For the students taking part in the task-based interviews, I anticipated that the change in their informal inferential reasoning would be significantly greater than that for all other students due to their additional work. These changes from pretest to posttest scores are displayed in Table 18 for the interviewees and the remaining 122 students.

Table 18

Change in Informal Statistical Inference Scores for Interviewees

	Interviewees $n = 14$ *M (SD)* % Increase	All Other Students $n = 122$ *M (SD)* % Increase	Mean Differences between Interviewees and All Other Students	Cohen's *d*
ISI	1.14 (1.35) 15.05%	0.99 (1.94) 13.29%	0.15	0.09
COMP	0.50 (0.94) 17.92%	0.17 (1.10) 5.61%	0.33	NA
PROB	0.29 (0.91) 10.98%	0.61 (1.10) 26.09%	-0.32	0.32
SAMP	0.36 (1.28) 16.29%	0.21 (1.18) 10.38%	0.15	0.12

The 14 interviewees had an increase of 1.14 points (15.05%) in their overall informal statistical inference (ISI) scores compared to 0.99 points (13.29%) for all other students. This was comprised of an increase of 0.50 points (17.92%) compared to 0.17 points (5.61%) for all other students for the comparing distributions (COMP) subscores; an increase of 0.29 points (10.98%) compared to 0.61 points (26.09%) for all other students for the sampling and

estimating a probability (PROB) subscores; and an increase of 0.36 points (16.29%) compared to 0.21 points (10.38%) for all other students for the inference based on a sample (SAMP) subscores.

The distributions for the change in informal statistical inference (ISI) scores and the subscores of PROB and SAMP passed tests for normality. Therefore, non-pooled two-sample *t*-tests were used to determine if there were significant differences in changes in these scores for the interviewees. A Wilcoxon Rank-sum test was used for the change in COMP scores. No differences were detected for the interviewees. This was an indication that the extra tasks the interviewees took part in did not have an impact on their improvement in informal inferential reasoning from the pretest to the posttest. However, I offer a reminder that there were significant improvements in informal inferential reasoning and in each of its subcategories for the 136 students.

Relationship between informal and formal inferential reasoning for interviewees.

I anticipated that for students taking part in the task-based interviews there would be a stronger relationship between their informal and formal inferential reasoning in comparison to that for all other students due to their additional work during the interviews. As stated previously, the Post COMP and Post PROB distributions did not pass tests for normality; therefore, Spearman's Rank Correlations were used. The results are displayed in Table 19 for the interviewees and the remaining students for comparison.

Table 19

Spearman's Rank Correlations between Informal and Formal Inferential Reasoning for Interviewees and All Other Students

	Post FSI	
	Interviewees (n = 14)	Non-interviewees (n = 122)
Post ISI	.21	.21*
Post COMP	.07	.06
Post PROB	.23	.21*
Post SAMP	.22	.15

*$p < .05$

No significant relationship was detected between the interviewees' informal and formal statistical inference scores. However, these correlations were at least as large as the correlations that existed for the remaining 122 students which were the same as those for the 136 students.

This group of 14 interviewees did not differ from the other 122 students in their initial informal inferential reasoning based on the pretest scores. The 136 students improved significantly in their informal inferential reasoning from the pretest to the posttest. These interviewees also did not differ in their improvement in informal inferential reasoning in comparison to the other students. This was an indication that the informal inferential reasoning demonstrated by these 14 interviewees was representative of the 136 students in the study. At the culmination of the study, the 14 interviewees demonstrated that they were stronger formal inferential reasoners than the remaining 122 students. This provided the basis for the qualitative analysis in the next chapter in which I will report on the informal and formal inferential reasoning of these 14 students as they perform the particular tasks in the task-based interviews.

Summary

To summarize these quantitative results, I will return to my research questions to review how this data assisted in answering those questions. My research questions for this study were:

1) For students enrolled in an introductory statistics class:

a) Does their informal inferential reasoning develop?

b) If their informal inferential reasoning develops, what are the characteristics of this informal inferential reasoning as it develops?

2) What is the relationship between students' informal inferential reasoning and their formal inferential reasoning?

Development of students' informal inferential reasoning.

By comparing students' responses to the informal statistical inference (ISI) questions on both the pretest and the posttest, it was evident that students' informal inferential reasoning did develop due to their regular classroom instruction. They had significant gains in their overall informal statistical inference scores and in each of its subcategories. These subcategories offered insight into the characteristics of their informal inferential reasoning by revealing how students performed on three different types of informal inferential reasoning questions. It should be noted that since differences existed by high school in the quantitative analysis, the results reported for the 136 students were not based on a simple random sample of introductory statistics students.

The first subcategory, COMP, assessed students when drawing conclusions by comparing two distributions of data. Watson and Moritz (1999) found that students who were able to see several aspects of data sets working together as a whole were best poised to make inferences when comparing those data sets. The high school statistics students in my study exhibited a significant gain in their responses indicating that they improved in their ability to bring together the important aspects of data sets (i.e. center, spread, shape) in drawing an informal conclusion. However, when analyzed by high school, neither group of students showed individual significant gains which may have been due to a ceiling effect.

The second subcategory, PROB, assessed students when drawing conclusions based on sampling and estimating a probability. The 136 students in this study exhibited significant improvement in their reasoning about how sampling and probability are both used in drawing an informal conclusion. These significant improvements also existed for students at each high school individually. The 46 students who were strong informal inferential reasoners based on their pre and posttest scores also showed significant gains in this category.

The third subcategory, SAMP, assessed students when drawing conclusions by comparing data from a single sample to the sampling distribution of all such samples. Saldanha and Thompson (2002) found that even after instruction on the sampling distribution, students tended to compare the results from a sample to the distribution of the original population rather than to the sampling distribution. Making an informal inference by examining where a sample statistic is situated in comparison to all such samples in the sampling distribution has the potential to support students' informal inferential reasoning and is necessary for formal statistical inference. The 136 students in my study exhibited a significant gain in their responses indicating that they improved in their ability to draw an informal conclusion by comparing a sample of data to its related sampling distribution. When analyzed by high school, only Deerfield students showed significant gains. The 46 students who were strong informal inferential reasoners based on their pre and posttest scores also showed significant gains in this category.

Relationship between students' informal and formal inferential reasoning.

My second research question asks about the relationship between students' informal inferential reasoning and their formal inferential reasoning. In examining the relationship between students' informal and formal inferential reasoning, correlations revealed that for the students in the study there was a significant positive relationship between their posttest informal

statistical inferences (ISI) scores and their formal statistical inference (FSI) scores. This was an indication that the stronger students were in informal inferential reasoning, the stronger they were in formal inferential reasoning. However, when the informal statistical inference scores were broken down into their three subscores for this study, the only significant relationship that remained was that between the sampling and estimating a probability (PROB) subscore and the formal statistical inference score. This was an indication that the strongest link to students' formal inferential reasoning occurred in their informal inferential reasoning associated with drawing conclusions based on sampling and estimating a probability. Konold et al. (2011) theorized that giving students the opportunity to estimate the probability of an event with an unknown *p* that cannot be summarized with a theoretical probability (e.g., unlike the probability of obtaining a sum of seven when tossing two die) supports their informal inferential reasoning by providing a conceptual understanding of the uncertainty that exists and a level of confidence in their inferences. These results provide support for Konold et al.'s theory. Students in each of the high schools and the 46 strong informal inferential reasoners showed significant improvement in their PROB subscores from the pretest to the posttest. It is possible that the second classroom activity during which students tossed small plastic pigs, along with other probability activities they took part in during their introductory statistics classes, supported their informal inferential reasoning when responding to the sampling and estimating probability questions on the posttest. This then led to a strong correlation to their formal inferential reasoning.

For students demonstrating substantial improvement in their informal inferential reasoning by the end of their introductory statistics class, there were no positive correlations to their formal inferential reasoning as was expected. However, for students who were strong

informal inferential reasoners, there was a significant correlation between their overall informal statistical inferences scores and their formal statistical inference scores. This was largely due to the significant correlation between their inferring about a population based on a sample of data (Post SAMP) subscores and their formal statistical inference scores. This was an indication that students who began as and remained strong informal inferential reasoners, particularly those strong in inferring about a population based on a sample of data, were also strong formal inferential reasoners.

When these same correlations between informal statistical inference and formal statistical inference scores were analyzed for each high school, these significant correlations between students' informal inferential reasoning and their formal inferential reasoning only existed for the students at Rosemont High School. This is not surprising since they scored better than Deerfield students on the posttest in all but one of the subcategories, Post SAMP, and had less variability in their informal statistical inference scores and subscores. This was an indication that high school attended had an impact on the relationship between students' informal and formal inferential reasoning. This was further explored in the qualitative analysis.

For the 14 interviewees, there was no significant relationship between their informal and formal inferential reasoning; however, the correlations were at least as large as those for the remaining 122 students. On the posttests, the 14 interviewees demonstrated that they were stronger overall inferential reasoners and stronger formal inferential reasoners than the other 122 students. With the qualitative analysis, I analyzed the task-based interviews with these 14 students to expand upon the answers to my research questions. I sought to better understand the development of students' informal inferential reasoning and its characteristics as well as the relationship between students informal and formal inferential reasoning.

Chapter 5 - Qualitative Analysis

The qualitative analysis reported in this chapter focused on the results of the four task-based interviews with the seven pairs of students who took part in them throughout their study of introductory statistics. I was looking for evidence of the development of students' informal inferential reasoning and for how their informal inferential reasoning related to their formal inferential reasoning. The analysis of the task-based interviews resulted in three main findings and a fourth that is situated within the third main finding.

The first finding materialized during the first task-based interviews as students compared distributions. The majority of the students recognized the differences in variability while comparing two distributions, but then subsequently applied only measures of center when drawing their informal conclusions. The second finding emerged during the third task-based interview when students were drawing an informal conclusion by comparing a sample of data to the corresponding sampling distribution. Students had exhibited an understanding of the sampling distribution, its characteristics, and the effects of sample size on its variability. However, when asked to take a sample to draw an informal conclusion, they did not want to rely on a single sample and they did not use the probabilities associated with the normality of the sampling distribution. The third finding came from the fourth task-based interview when students exhibited procedural knowledge of formal statistical inference but limited conceptual knowledge. Most of the students could not accurately state what the percentage associated with a confidence interval meant and could not accurately define the *p*-value when conducting a hypothesis test. Following the discussion of each finding, I will return to the interviewees' responses to the corresponding informal statistical inference questions on the pre and posttests to discuss consistencies and inconsistencies in their informal and/or formal inferential reasoning. The fourth and final finding emerged with the exploration of a quantitative result indicating that

the relationship between students' informal and formal inferential reasoning on the posttests differed by high school. Interviewees' work during the last task-based interview provided some insight into this quantitative result. All task-based interview questions and pre/posttest questions can be found in the appendices.

Reliance on the Mean or Median When Comparing Distributions

While comparing distributions during the first task-based interview, the majority of the interviewees relied upon measures of center to draw conclusions based on the histograms they were shown. This occurred even when the variability was an additional factor to be contended with in making a decision. However, the students did recognize the variability as they alluded to it in a variety of ways.

There were five parts to the first task-based interview, each comparing data of children's test scores from two classes. The student pairs were asked if the classes scored equally well or if one of the classes scored better. The first part required a comparison of the measure of center with one of the classes clearly scoring better. The second part also required a comparison of measures of center, however, the students had to take the shape (one was skewed right and one was skewed left) into consideration. The third part added complexity as students were shown two distributions with the same mean but different variability. With yet another layer of complexity, the fourth part showed two classes of different size in which students had to reason proportionally in determining which class performed better. To complete the fifth and final comparison, students needed to combine their proportional reasoning with the concept of variability.

In Part 1 of the task-based interviews (Appendix B) on comparing distributions, all students inferred correctly that the Red class scored better than the Blue class based on the data

in the histograms when the inference required only a comparison of means and/or medians. Six of the seven student pairs also inferred correctly in Part 2 that the Green class scored better than the Purple class when the comparison required recognition of the effects of skewness on the means and/or medians. In Part 4, five of the student pairs inferred correctly that the Black class with a left-skewed distribution scored better overall than a larger class with a symmetrical distribution. When the comparison required students to compare the measures of center or the effects of skewness on the measure of center, the majority of the pairs reasoned correctly about which class scored better overall.

The two remaining comparisons, in Part 3 and Part 5 of this task, required students to take the variability into consideration along with the measure of center. In Part 3, the interviewees were shown the two symmetrical distributions of the same size in Figure 1.

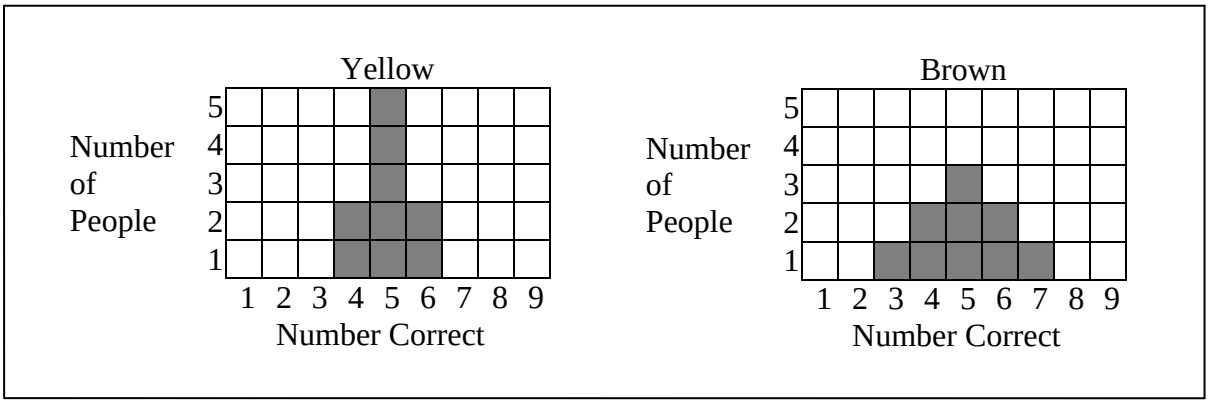


Figure 1. Comparing distributions Part 3. Adapted from "The Beginning of Statistical Inference: Comparing Two Data Sets," by J. Watson and J. Moritz, 1999, *Educational Studies in Mathematics, 37*(2), p. 151. Copyright 1999 by Springer.

I was looking for evidence that students would use the variability as the basis for deciding which class scored better when the mean/median was not the distinguishing factor. In that case, the distribution with the smaller variation would be considered the better-scoring class. Four pairs and one student from a fifth pair concluded that the classes scored equally well. These nine students were basing their decisions on the mean and/or the median which was the same for

both the Yellow and the Brown classes. The variety of ways in which they expressed their reasonings is included in Table 20.

Table 20

Students' Responses and Remarks in Part 3 of First Task-Based Interview

Students (conclusion)	Students' Remarks
April (equal)	I would say like, it [the two classes] would [have scored equally] just because if you put like the two boxes at the three and seven on top [from the Brown class], they'd be the exact same thing [as the Yellow class]. So I feel like it would just even out cause one's higher than the five and one's lower than the five.
Brian (equal)	Yeah, I would say in this case the median and the average are the exact same thing.
Caitlin (Yellow to undecided)	Um, I'm going to say Yellow. Probably because the five has the most [in the Yellow class] and this [the Brown class] is more spread out.
David (equal)	Well I say that the Brown, there's a three and a seven and they pretty much equal each other out. So you could add that to the five.
Fritz (equal)	You'd have to look at the entire picture of the Brown class as well cause there's a seven there and if you do like each box. They cancel out or to add out, it's pretty much equal.
Emily (equal)	I have to look at the fives that the Yellow class had that the Brown class didn't have. I mean, yeah, you have a seven [in the Brown class] but there's also five fives in that class [the Yellow class]. Where we only have three in the Brown class. And we had an equal amount of sixes too.
Gabrielle (Yellow)	Like the three and the seven [in the Brown class]. Like if you were to have a boxplot, this would be the low [three], this would be the high [seven] and that would be the middle [five]. Like this one's more closer together [the Yellow class].
Jared (equal to Yellow)	I guess what I'm thinking is that statistically, they're the same. They scored the same but by the looks of the graph, it's starting to look like Yellow did better. Just because it doesn't have, they're not really outliers but, they kind of look like outliers throwing it off.

Laura (equal)	And this one if you took the, I think if you took the top two numbers [fives] off of the Yellow graph, and put them on each side, you'd have the same exact one [same graph as Brown].
Mark (equal)	Yeah, I'd say the average is probably the same like exactly the same because they have the same amount revolving around five.
Steve (equal)	I think they scored equally well because what I used as a frame of reference, I took this and these two numbers [three and seven in Brown class] which were the, not the outliers, but the farthest ranges of each data and I put them in the middle. So if you did that they [the Yellow and Brown graphs] look exactly the same. So, furthermore, these two can just cancel each other out if you want to look at it like just ball parking what the averages are. So that if you just like eliminate each end, you just have five. So it would seem that each has the same average of five.
Rachel (equal)	Just the fact that it's, I don't know how to explain this, but like it just seems more equal since this is just, since the highest is five and then like what Steve said, these two [three and seven in Brown class] cancel each other out so therefore it would still be five. Like the highest and the mean, five.
Nathan (undecided)	It [graph of Brown class] has the highest but it also has the lowest. This [the Yellow class] has a lot more fives than this one [the Brown class].
Pete (undecided)	Yeah, this one [the Brown class] is more variable which gives us the graph here. Whereas this has the higher number of people [scoring five] which means it's more accurate, kind of. Whereas, like this class [Brown] could be a lot of smart people and a lot of people who didn't do so well. Where this one's [Yellow] kind of like everyone's the same.

All of the students deciding that the classes scored equally well referred to the two scores of three and seven correct in the Brown class as equivalent to the two additional scores of five correct in the Yellow class or that the three and seven correct in the Brown class could be placed in the center at five correct and the two distributions would be identical. Even when questioned about the greater number of students scoring five in the Yellow class, these students did not change their conclusions. This discussion with Brian and April represents a typical response.

Brian: I'd say equally as well.

April: I would say like, it [the two classes] would [have scored equally] just because if you put like the two boxes at the three and seven on top [from the Brown class], they'd be the exact same thing [as the Yellow class]. So I feel like it would just even out cause one's higher than the five and one's lower than the five.

Brian: Yeah, I would say in this case the median and the average are the exact same thing.

April: Cause they're both normal distributions too.

Brian: Yeah, they're both normal distributions.

Interviewer: How about if someone else said, “I think the Yellow class scored better because they have more people at that mean or median value so I think the Yellow class scored better." How would you respond to that?

April: I would say they might have had more at the five but the Brown people have someone that was higher than the five and some people that were lower.

Brian: I'd say, find the quartiles of each, they'd probably be about the same so the Yellow couldn't have done better.

April and Brian refer to these distributions as normal distributions although symmetrical would be more accurate. Even when questioned about the Yellow class having more data at the mean or median value, they argue that this does not change the fact that the mean or median values are the same. Brian also justifies this by saying that the quartiles will be about the same. The reality is that this variability would also be evident when comparing quartiles. The concept of symmetry or balance in these two distributions outweighs the differences in variability they see in the

distributions. April and Brian do not ignore the variability but they are convinced that it does not indicate that the Yellow class scored better than the Brown class.

Laura and Mark, another pair who concluded the classes scored equally well, referred to the consistency of the Yellow class data.

Mark: I would say they score equally but Yellow is more consistent.

Laura: Yeah, equal averages but Yellow is more consistent with the higher numbers. So that's something like, a point that could be very argued because like the seven [from Brown class] could be taken with that too so it evens out. So I'd say consistent.

Interviewer: Exactly what do you mean by consistent?

Mark: Towards the same…

Laura: Mean.

Mark: Around yeah around, less variation.

Laura and Mark were reasoning that the Yellow class had more fives and was more consistent but the seven in the Brown class could be argued to be equivalent to the greater number of fives in the Yellow class. For this pair, exactly what constitutes scoring "better" may not have been clear. Laura and Mark reasoned in much the same way that April and Brian did in determining that the classes scored equally well. Although, Laura and Mark's comments do reveal that they find it necessary to add the caveat about the smaller variability in the Yellow class data.

The students in this next pair are not in agreement in drawing a conclusion. Caitlin demonstrated global reasoning pertaining to the differences in variability by inferring correctly

that the Yellow class, with its smaller variability, scored better overall. However, the frailty of her conclusion can be seen in her conversation with her partner David.

David: I say equally.

Interviewer: Equally?

David: Cause I think they are both pretty normal distributions and they each center at five. Equal's good.

Caitlin: Um, I'm going to say Yellow. Probably because the five has the most [in the Yellow class], this [the Brown class] had less fives than this and this [the Brown class] is more spread out. And the three is going to bring it down quicker too. The kind of like the median of that.

Interviewer: So what do you mean the three is going to bring it down? Bring what down?

Caitlin: Like the average, maybe, that's what it was. I'm trying to, I can't really see how it's equal. I'm still going to go with Yellow.

Interviewer: So how do you argue [asking David]?

David: Well I say that the Brown, there's a three and a seven and they pretty much equal each other out. So you could add that to the five.

Caitlin: And these are the same, the four and the six and then the five even though that's still higher, they're still the largest median? Ok, yea, I can kind of see it.

Interviewer: But what about the argument that there are more fives here in the Yellow class [asking both Caitlin and David]?

David: That doesn't change the average, but the three and the seven would make two fives so they average exactly.

Caitlin: Ok.

Interviewer: So what would you go with?

Caitlin: I think I would go with equal again. I know, thanks for that [talking to David]. Yeah, cause I see the ten and if there were to be two more fives.

Interviewer: So you don't think that the spread matters?

Caitlin: I mean it does but since there is a higher one on this side, it's just going to equal, it's going to be more than the three. It's going to make it higher anyways. Unless there was since that seven that just cancelled the three out and there would be four which isn't as much as this one would still be. It's making me kind of confused still.

Interviewer: So David, you're pretty convinced [the classes scored equally well] and Caitlin, you're looking at it differently.

Caitlin: Yeah. I mean you could add and then that's how he got the two fives. Unless since this was higher, it would cancel that out. If you left that a little lower average it wouldn't equal ten again to get the same as that.

David: If you cancel them it would be.

Caitlin: Right. I can't conclude really.

Caitlin is influenced by her partner who is convinced that the classes scored equally well. She tried to put David's reasoning into her own words but still could not ignore the differences in variability in making her decision.

This pair decided that Pete would find the average score for the Yellow class and Nathan would find the average score for the Brown class with their calculators. Pete notices the greater variability in the Brown class histogram; and after realizing the averages are the same, they discuss further the differences in variability between the two classes.

Pete: There's more variability here [pointing to the Brown class histogram], I think these are the…

Nathan: Yeah, that's what I was thinking. I got, what'd you get for the mean?

Pete: I got, um, five.

Nathan: Aw, it's the same. So they're both five.

Pete: But when I first looked at it, I thought it was this one [pointing to the Brown class histogram] because of the number of people. It's more, it has a bigger range here, there's more.

Nathan: It has the highest but it also has the lowest.

Pete: Right, the more variability here.

Nathan: This has a lot more fives than this one.

Pete: Yeah, this one is more variable which gives us the graph here [pointing to histogram of Brown class data]. Whereas this has the higher number of people [scoring five in Yellow class] which means it's more accurate, kind of. Whereas, like this class could be a lot of smart people and a lot of people who didn't do so well [referring to Brown class]. Where this one's kind of like everyone's the same [referring to Yellow class].

They are seeing the difference in variability between the scores for the Yellow and the Brown classes; however, they did not come to a conclusion about whether the two classes scored equally

well or one scored better than the other. Both Nathan and Pete reasoned similar to Caitlin and even though they agreed, they found it difficult to come to a conclusion.

The only pair to conclude that the Yellow class scored better than the Brown class was Jared and Gabrielle. Initially, Jared stated that the classes scored equally well but subsequently changed his mind when questioned further and after listening to Gabrielle's comments.

Jared: I'm thinking that they're about the same again average-wise because they're like both around the same score. There's no differences. They're the same average I can tell by seeing but just different scores, if that made any sense at all.

Interviewer: Gabrielle, what about you, what do you think?

Gabrielle: I'm not, I don't know. I want to say Yellow but I don't know why. Just because it's like more clumped together and you can obviously see the middle. But on here [the Brown class histogram] five is the middle too, so like I'm not sure.

Jared: The images are like symmetrical at the same spot and that's where I'm assuming the average would be like a 4.5, so…

Interviewer: Let me ask you this question. How about if someone came in and said, well, I think that the Yellow class scored better because they have more fives than the Brown class? What would you say to that? Would you agree, disagree?

Jared: I could see where they're coming from. That Brown does have that seven there, that individual scored higher but there's also an individual that

scored three, so I think that Yellow did, I guess you could say that Yellow did do a little bit better cause, I don't know, like, I know how to call…

Gabrielle: They don't have that extra person like throwing them off, I guess.

Interviewer: Which extra person do you mean?

Gabrielle: Like the three and the seven [in the Brown class]. Like if you were to have a boxplot, this would be the low [three], this would be the high [seven] and that would be the middle [five]. Like this one's more closer together [the Yellow class].

Interviewer: So does that weigh in when you're deciding which one scored better or if they scored the same?

Jared: I guess what I'm thinking is that statistically, they're the same. They scored the same but by the looks of the graph, it's starting to look like Yellow did better. Just because it doesn't have, they're not really outliers but, they kind of look like outliers throwing it off.

I interpreted Jared's comment that statistically the classes were the same to mean that their averages and/or medians were the same. Both Jared and Gabrielle recognized that the measure of center did not completely determine which class scored better. Without specifically using the term variability, they were clearly indicating that the smaller variability in the Yellow class data was the reason they chose the Yellow class as scoring better than the Brown class.

Evidence of students' informal inferential reasoning was determined by their abilities to (1) make an inference based on the data, (2) use the data as evidence for their inference, and (3) use probabilistic language to indicate a level of certainty in their inference (Makar & Rubin, 2009). All but three students were able to make an inference based on the data. However, only

two students from a single pair inferred accurately based on the differences in variability between the Yellow and the Brown class data. The three students who did not did make an inference and remained undecided about which class scored better or if the classes scored equally well, were struggling with how to interpret the differences in variability between the two classes. These students' informal inferential reasoning had actually developed further than the majority of interviewees who responded that the classes scored equally well. Those inferring that the classes scored equally well were taking only the measure of center into consideration. All students used the data as evidence for making their inference. Again, it was a matter of how they used that data and to what extent they considered all aspects of the data. The probabilistic language used by all of the students came in the form of hedging their inferences. Terms like "probably" and phrases like "I'm thinking" or "I would say" indicate that they were unwilling to make a definitive statement.

Similar to the interviewees' responses in Part 3, when the distributions were symmetrical but of differing sizes in Part 5, six student pairs again recognized and commented on this variation but did not take the variation into consideration when drawing their conclusions. The distributions are displayed in Figure 2.

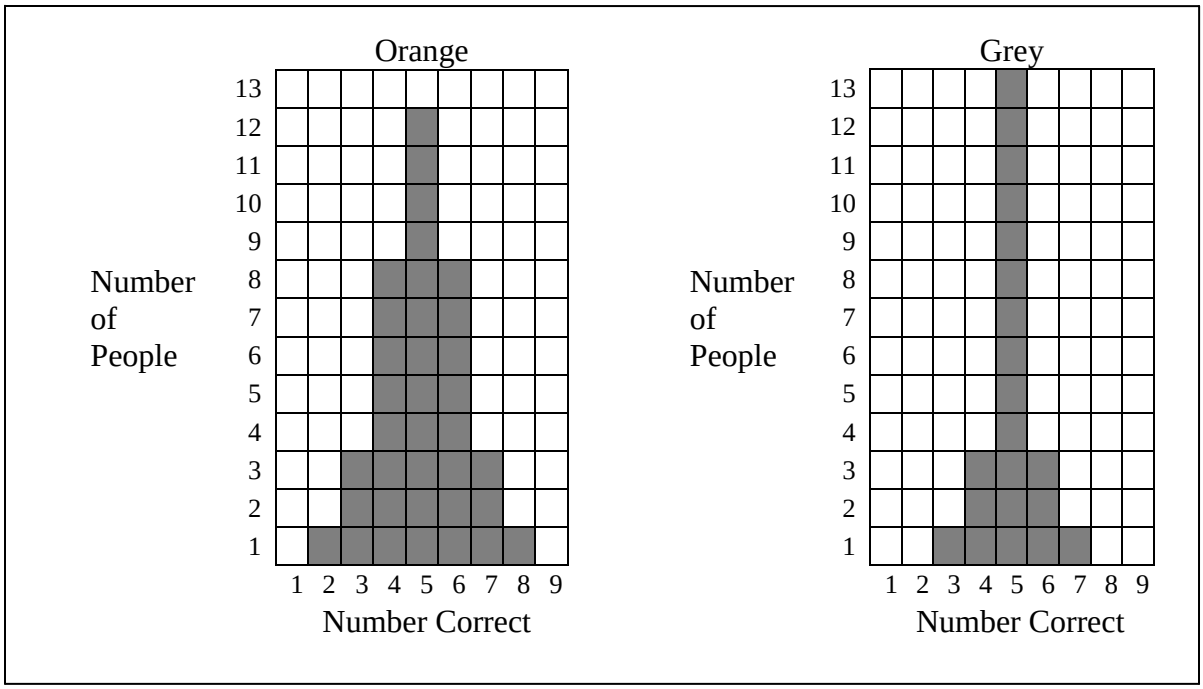


Figure 2. Comparing distributions Part 5. Adapted from "The Beginning of Statistical Inference: Comparing Two Data Sets," by J. Watson and J. Moritz, 1999, *Educational Studies in Mathematics, 37*(2), p. 151. Copyright 1999 by Springer.

Table 21

Students' Responses and Remarks in Part 5 of First Task-Based Interview

Student Pairs	Students' Remarks
Brian (equal)	Oh, what I mean is between each standard deviation it would probably be approximately, it's kind of hard to explain what I'm thinking but basically the standard deviation, the distance between one standard deviation to the next between the Grey class and the Orange class will probably be exactly the same because even though the Orange has more people the curve is the same, the mean is the same, the average is the same, and the quartiles are also the same.
April (equal)	It's basically like the same shape except just the Orange has more, if you were draw the line of the distribution around it. That's why I would think that it would be the same. I don't think it's so much of a difference that you could say that they [the Orange class] had lesser score than the other one [the Grey class].
David (equal)	I'd say they scored equal because they both have normal distributions. Even though the ranges are different they're still symmetrical on both sides of the numbers.

Caitlin (equal)	Um, yeah, I'm going to say they're equal too cause of the symmetrical distribution. And even though this one [the Grey class] has another person [scoring five], close to this one it still adds on to it and makes the average higher.
Emily (Orange)	Yeah, just first glance I would say Orange class but there's more to it. Cause when you look at the five, in both of them, there's only one more five in the Grey class than the Orange class. But you also have a good, decent amount of fours and sixes where when you look at the Grey class you don't really have that. But also you have to look at there's a larger amount of people in the Orange class than the Grey class.
Fritz (equal)	They're both almost, pretty symmetric. And the values like the fives are almost the same but there's like a lot more [in the Orange class], there's an eight there too and seven sixes. I think equal just because of the symmetry cause like the two and the eight there [in the Orange class]. It's almost the same, it's just like that [the Grey class]. I think it's the same.
Jared (Orange)	I'd go with the Orange. Again it might come down to what if there were more people with Grey but I can't solely base it off of that. So with what we have, I would go with Orange.
Gabrielle (Orange)	Yeah, there's more people scoring more numbers correct.
Laura (equal)	Um, same kind of thing where like, well, actually, no I'd say equal cause I think that like looking at the data the five being like the median, like the five's the middle high point of the graph. And um, it is on both of them. But then if you look at the two sides of the graph, they don't have the same amount of data. They have the same exact amount on each side of the five so I think the five would be the mean for both of them, with looking at that.
Mark (equal)	I think they would be pretty even cause, you know, the Orange has more, I would say, smarter, they also have less smart.
Steve (equal)	I'd say they're probably like about equal just because it's symmetrical. Yeah and this [the Grey class] is a smaller class size, so like, you don't know if you have more people it would look more like this [the Orange class].
Rachel (Orange)	Yeah, and if you take into account that if you did interview more people their average would most likely be five. So then I don't know if that would make, it would make [the average of] this data set [the Grey class] larger than this one [the Orange class] but I think, I still think [the average for] the Orange data set would be larger even if more people did get interviewed. They would most likely get five.

Nathan (Grey)	Like there's, it's more dense so with this one [the Orange class] the outliers really don't affect it too much cause there's so many numbers. But this one [the Grey class] the outlier would probably affect it. It's closer to here [the center of five] so probably wouldn't it, like a three could have an effect on it. But this two [outcome in Orange class] probably wouldn't have as much affect because of the bigger number of people in the [Orange] class.
Pete (Grey)	I think every one we've interpreted we're just taking it on pure mathematical, the mean, which can be unfair. But in the histograms, we're not because these are both pretty symmetrical data. But we're taking it based on mathematical data and the fact of the sample size. So that's the two things we were determining and variability also.

Jared and Gabrielle find the similarities between the comparison of the Orange and the Grey classes in this last part of the activity to the comparison of the Yellow and Brown classes in Part 3. However, they are inconsistent in their reasoning between the two parts. The added complexity of differing sample sizes had Jared suggesting that the increased variability in the Orange class data meant that the Orange class scored better.

Jared: This is kind of like the other one, the, which one was that?

Gabrielle: That one [pointing to Part 3].

Jared: Yeah the third one. It's almost the same concept. Like Brown is Orange and Yellow is Grey. 'Cause it's got more up the middle and this one is more compact [referring to the Grey class]. This time the Orange one has more people in it so the "what if" factor comes in again. But, my first reaction I guess I would say Orange is better even though it doesn't have as many people scoring the middle there, it's got other people that are right there. Like there's eight people at four and six and this one [Grey class] there's only three people at 4 and 6. I'd go with the Orange. Again it might

come down to what if there were more people with Grey but you can't solely base it off of that. So with what we have, I would go with Orange.

Interviewer: You would agree, Gabrielle?

Gabrielle: Yeah, there's more people scoring more numbers correct and getting...

Jared: Oh, yeah, I didn't even look at that, there's a bigger range [in Orange class histogram].

Interviewer: So let me ask you this question. What if somebody said, "Yeah they do have some more sevens and eights over here [in Orange class data] but they also have more twos and threes and this Grey class has more that scored that five in the middle." How would you respond?

Jared: Again I think I would go with the next level, the next step outward.

Interviewer: You're saying from the five.

Jared: From the five, there's more fours and sixes in the Orange. Instead of, yeah, there's a couple twos and threes and a couple sevens and eights. This one doesn't have as many fives but it's got more sevens than the Grey, more sevens and eights.

The differences in class sizes added more complexity as it also required thinking proportionally. Jared pointed out the increased variability in the Orange class data but did not use it proportionally in making his decision. The larger data set for the Orange class with more scores above the center appeared to be the most important piece for Jared.

Two additional students from different pairs reasoned as Jared did about the larger amount of fours and sixes correct in the Orange class as indicating that the Orange class scored better. Both of their partners decided the classes scored equally well based on the symmetrical

shapes of the distributions, as did three other pairs of interviewees. This response from Caitlin and David represents a typical response for this reasoning.

David: I'd say they scored equal because they both have normal distributions. Even though the ranges are different they're still symmetrical on both sides of the numbers.

Catlin: Um, yeah, I'm going to say they're equal too cause of the symmetrical distribution. And even though this one [the Grey class] has another person [with a score of five] it's close to this one [the Orange class] and still adds on to it and makes the average higher, and then the extra ones on these [referring to Orange class].

David: The averages are going to be the same.

Interviewer: So how about if somebody said, "But there are more students who scored seven here and a student who scored an eight so I think the Orange class scored better." What would you say? How would you respond to that?

Caitlin: I'd just explain that even though, you said these two, seven and eight? Um, they're not really outliers but there's a bigger, there's like, the mode is kind of in the middle here and even though there's more of them, it's going to bring it up anyway. So that's how I'd explain it.

David: Yeah, I'd say the sample size again. They've got the same average; they've got the same shape. Then you know it's got more on the two and three side too; the same argument that way.

Interviewer: How about if somebody said, you had mentioned this Caitlin, if somebody said, "There's an extra person who scored a five here and this is not, the

range isn't as big, so I think that the Grey scored better." What would you say to someone who said that?

Caitlin: I guess I would just say that since it's another person in the median of the average it could be, the other person could be with a lower average which would bring it down anyways. So that person added, kind of brought it up.

Interviewer: Anything else?

David: Not really. The shapes are what keeps me thinking that they are equal. As long as you're saying score-wise, an average, they're going to have the same. Mathematically they're equal. The numbers are.

David's response was consistent with his response in Part 3. He focused on the symmetrical shape of the distributions and the measures of center in concluding the classes scored equally well. Caitlin, who was struggling with how to interpret the differences in variability in Part 3, was unable to reason proportionally in terms of the variability. The shape of the distributions and the larger sample size in the Orange class were the features she focused on in Part 5.

The only pair to conclude that the Grey class scored better was Pete and Nathan. As they had done previously in Part 3, each used their calculators to find the averages first before answering the question posed.

Pete: I got like 5.13.

Nathan: I got five.

Pete: So again, I think that would say that, in general, the Grey class did better because it has a pretty comparable mean even though the Orange class had more students. I think if more students were added to the Grey class it would make the Orange class's more. Or if more students were added to

the Grey class it would increase the mean versus the Orange class's because of more.

Nathan: So like, there's nobody that has two.

Pete: And the minimum and maximum.

Nathan: Like there's, it's more dense so with this one [the Orange class] the outliers really don't affect it too much cause there's so many numbers. But this one outlier would probably affect it. It's um closer to here so probably wouldn't, it, like a three could have an effect on it [the Grey class]. But this two probably wouldn't have as much affect because of the bigger number [in the Orange class].

Interviewer: The bigger number?

Pete: Of people in the class.

Interviewer: So what is it about the, between the maximum and the minimum, there. How does that play into your decision?

Pete: Well I just think it's just because the number in the class. This one's, this one's bigger so the smaller and the higher numbers won't have as much affect cause there's so many more numbers to cancel them out. As opposed to this three probably has more effect than this two would because there's a smaller number. And there aren't as many numbers that would cancel out that outlier.

Interviewer: So which one of those classes again do you think?

Nathan: I would say the Grey class. Because with more people it would probably end up making the mean a lot different. It would be able to cancel out the three and maybe the seven too and like help raise it.

Pete: I think every one we've interpreted we're just taking it on pure mathematical, the mean, which can be an unfair but in the histograms we're not, because these are both pretty symmetrical data. But we're taking it based on mathematical data and the fact that the sample size so that's the two things we were determining and variability also.

Interviewer: And how does the variability play in here?

Pete: Variability wouldn't really apply as much because…

Interviewer: First of all, what do you mean by variability?

Pete: When I mean variability, I mean the minimum and the maximum and the distance between and how far those values are taking the data. So, as Nathan was saying, the three and the seven, how those two. Whereas if we had one that was, got rid of the three and the seven, it was just four, four, four, all fives then six, six, six. That would have little variability cause it's going from four to six. So if you saw that like on a boxplot, it would be very tight.

Interviewer: So how does that affect your decision? How's that variability affect your decision when you're choosing one class over another?

Pete: It affects us cause it's um, it's telling us how much our numerical data that we're taking or judging it on, how much that's going to be affected because of those outliers or technically not in this situation but. Cause

> most of this has been pretty symmetrical but it still needs to be taken into account the people who did not do very well and the people who did extremely well.

Nathan and Pete were unsure about the effects of the differing variabilities in Part 3, but the differing sample sizes in Part 5 appeared to clarify their thinking. The consistency in the Grey class data was their focus in deciding that the Grey class scored better than the Orange class.

Returning again to the Makar and Rubin (2009) framework for thinking about informal statistical inference, all students were able to make an inference based on the data. However, only two students from a single pair inferred accurately based on the differences in variability and sample size between the Grey and the Orange class data. All students used the data as evidence for making their inference with the majority of them referring to the symmetrical shapes of the distributions. Again, it was a matter of how they used that data and to what extent they considered all aspects of the data. The probabilistic language used by all of the students in Part 5 was similar to that in Part 3 when students hedged in stating their inferences.

Students' responses in Part 3 and Part 5 of the first task-based interviews indicated that when students had to draw a conclusion based on the variation, they saw the differences in variability between the two distributions but the majority of them did not use that in concluding which class had scored better. When comparing symmetrical distributions with the same mean and median, the equitable measures of center was more important than the differences in variability to students in drawing a conclusion. I then returned to the interviewees' responses to questions 1 through 4 on the pre/posttests where they compared distributions of data displayed in dot plots and boxplots.

Pre/posttests.

The interviewees had little difficulty in answering the first two questions about interpreting the results of an old and new medication for headache relief. The data on minutes to relief were displayed in the dot plots in Figure 3. The students were presented with two statements, one of which argued that the old medication was more effective and the other that the new medication was more effective.

On the pretest, twelve of the fourteen interviewees correctly responded "invalid" to Question 1 that the old medication was more effective. Thirteen of them correctly responded "valid" to Question 2 with the statement that the new medication was more effective. Only one of the students responded incorrectly to both of these questions.

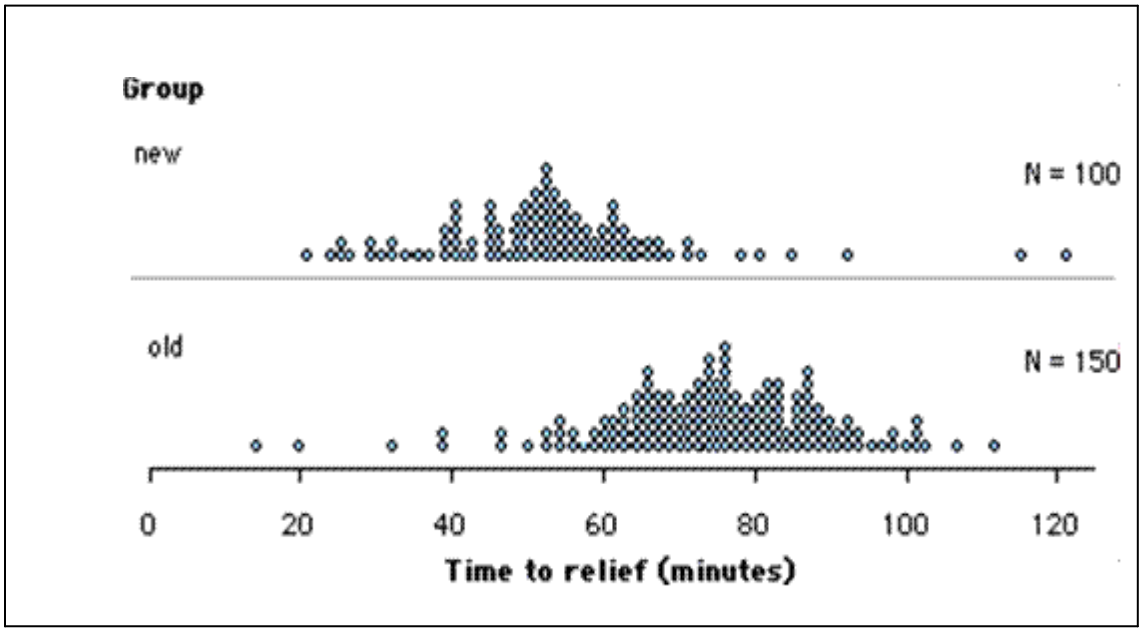


Figure 3. Comparing distributions questions 1 and 2 on pre/posttest. Adapted from "Comprehensive Assessment of Outcomes for a first course in Statistics (CAOS)," developed by the Web ARTIST Project, Principal Investigators J. Garfield, R. delMas, B. Chance, and A. Ooms, 2005.

On the posttest, all fourteen interviewees responded correctly to question one and only one student (different student than on the pretest) responded incorrectly to question two. These results are summarized in Table 22.

Table 22

Pretest/Posttest Results of Comparing Distributions Questions 1 and 2

Pretest/Posttest response	Number of students - Question 1	Number of students - Question 2
Correct/Correct	12	12
Incorrect/ Correct	2	1
Incorrect/Incorrect	0	0
Correct/Incorrect	0	1

Answering these two questions required students primarily to compare measures of center. Some students may also have reasoned proportionally as the data sets were not of the same size.

The second two questions proved to be more difficult for some of the interviewees. Students were shown four sets of boxplots comparing running times of athletes split according to those who were in a weight training program and those who were not weight training (Figure 4). Students were first asked to identify the set of boxplots providing the most convincing evidence that the weight training program was effective in decreasing the athletes' running times (choice D). They were then asked to identify the set of boxplots providing the least convincing evidence (choice B). These two questions required students to compare medians and to compare variability. These results are summarized in Table 23.

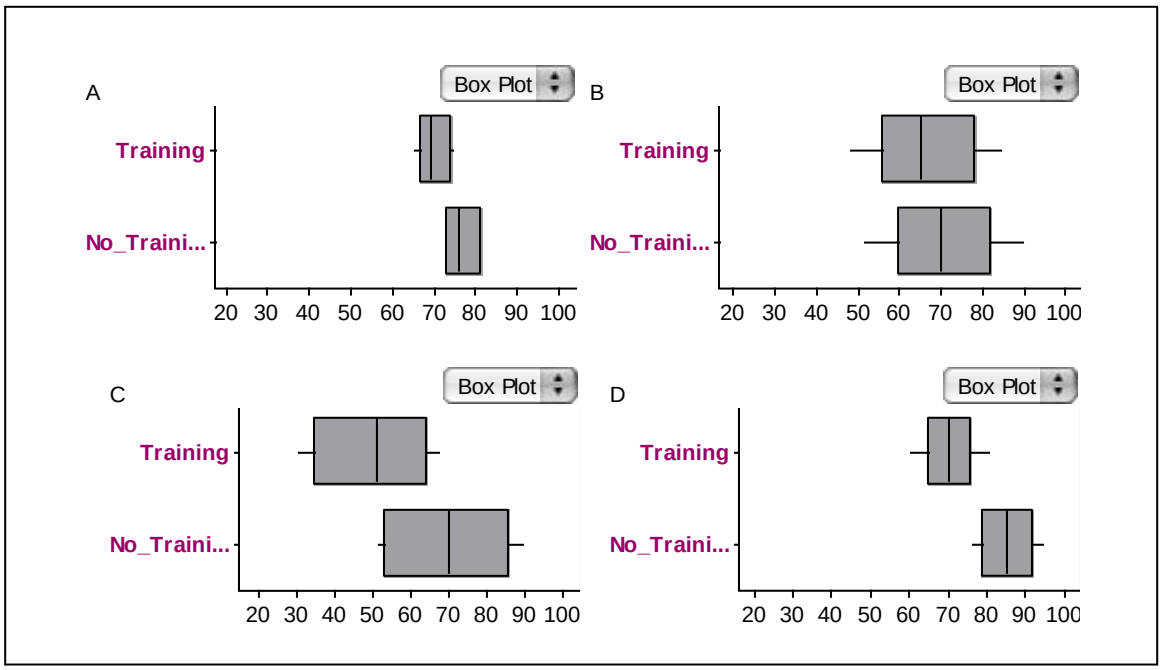


Figure 4. Comparing distributions questions 3 and 4 on pre/posttest. Adapted from "A Framework to Support Research on Informal Inferential Reasoning," by A. Zieffler, J. Garfield, R. delMas, and C. Reading, 2008, *Statistics Education Research Journal*, *7*(2), p. 49.

For each of these questions, half of the interviewees answered them correctly on the pretest. For Question 3, the number of interviewees answering correctly decreased from seven to six. Seven of the eight students answering incorrectly about the most convincing evidence chose the set of boxplots in C rather than D. These students recognized that a large difference in medians would be most convincing but failed to recognize that a smaller variability would be most convincing. Students showed improvement on the posttest for Question 4 about the set of boxplots in B providing the least convincing evidence. All but one interviewee answered correctly. This may have been an indication that they recognized the small difference in medians and the large overlap, thus larger variability, in the boxplots.

Table 23

Pretest/Posttest Results of Comparing Distributions Questions 3 and 4

Pretest/Posttest response	Question 3 – Most Convincing Number of students	Question 4 – Least Convincing Number of students
Correct/Correct	4	7
Incorrect/ Correct	2	6
Incorrect/Incorrect	5	1
Correct/Incorrect	3	0

The possibility exists, however, that the interviewees did not take the variability into account in either Question 3 or 4. These two sets of boxplots, B and C, look very similar with the largest variability of the four. The main difference in the two sets is the difference in their medians. Here once again, comparing the measures of center, the medians, appears to have been the most important aspect in drawing a conclusion.

These results indicate that early in their study of introductory statistics, the concept of variability was a difficult one for students to use in drawing conclusions.

Instead, students often used only measures of center when comparing distributions of data. This occurred whether the data was presented in histograms as in the task-based interviews or boxplots on the pretest. During the task-based interviews, the majority of the interviewees acknowledged the variability but then did not take it into consideration as they compared class scores. It was possible that they did not know what to do with that information early on in their study of statistics. Their exposure to the limited amount of statistics required in previous high school courses included primarily measures of center. A global view of two class sets of test scores with equal measures of center and differing variability would, in most scenarios, deem the data set with the smallest variability as the better scoring class. Therefore, it was also possible that the majority of the interviewees simply did not believe that the more consistent set of test

scores warranted the generalization that the class scored better. However, this difficulty with variation persisted through to the posttest as students compared boxplots of running times. The majority of the interviewees continued to rely primarily on the median in making their comparisons.

Drawing Conclusions with the Sampling Distribution

The third task-based interview focused on the sampling distribution and drawing a conclusion from a sample of data. In the first two parts of the interview, the interviewees reviewed the characteristics of sampling distributions which they had recently studied in their statistics course. It was in the third and final part of the interview, when students were asked to draw a conclusion based on a sampling distribution, that they encountered difficulty.

Part 1 of the third task-based interview had students repeating the first portion of the classroom activity in which they predicted what sampling distributions would look like for a given population distribution and considered the variability of those sampling distributions. In the second part of the task-based interview, students viewed three sampling distributions generated from the Random Rectangles activity in *Fathom*. The sample size increased from five to 10 and then to 25 as the average area was graphed. Students are asked what average areas would be likely and which would be rare or unlikely based on each of the sampling distributions. The interview concluded by returning to the Tossing Monopoly Houses Activity from the second task-based interview to test the hypothesis that Monopoly hotels have the same probability as the houses of landing upright. Students were shown a sampling distribution of the proportion of Monopoly houses landing upright generated from 200 samples of size 10 to use in drawing their conclusion.

In Part 1 of the task-based interview, students were shown the population distribution with its mean, median, and standard deviation and the five sampling distributions displayed in Figure 5.

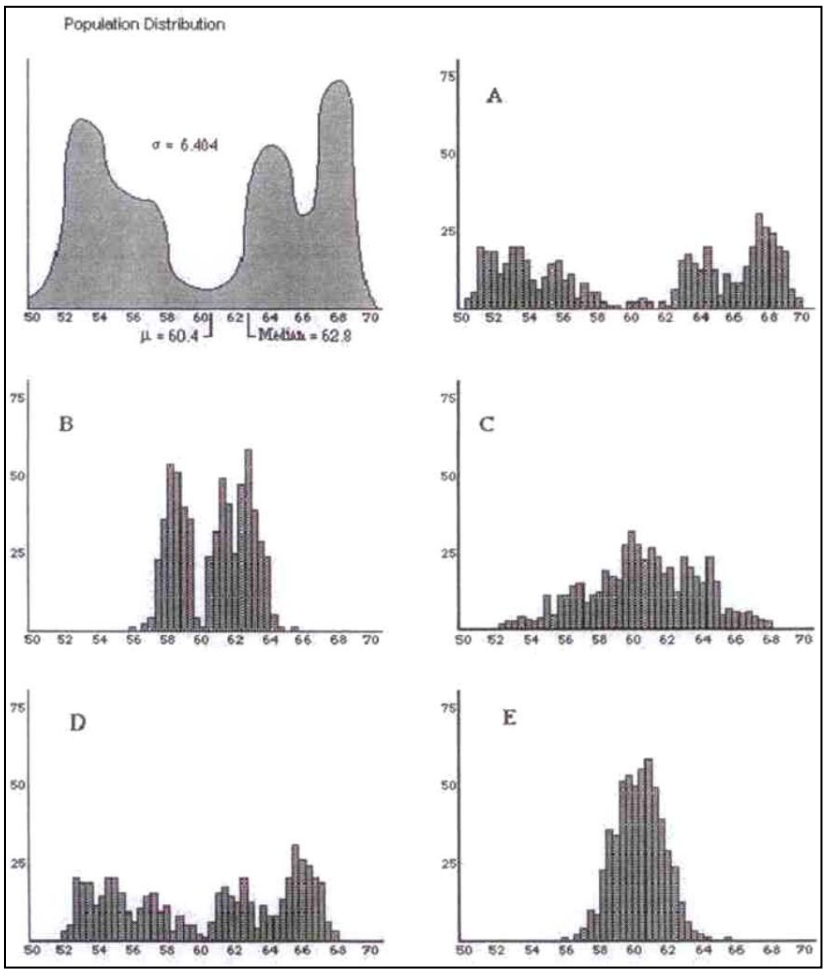


Figure 5. Sampling distributions Part 1. Adapted from "Reasoning about Sampling Distributions," by B. Chance, R. delMas, R., & J. Garfield, 2004, in D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, p. 321. Copyright 2004 by Kluwer Academic Publishers.

Students chose the distributions that represented 500 samples of size 4 and size 16 from the five possible sampling distributions. All seven pairs of interviewees chose sampling distribution graphs that were approximately normal in shape. Six of them also correctly identified the effect of sample size on the variability of the sampling distributions. Only one pair incorrectly

identified the variability of the sampling distribution for a sample of size four; however, they did correctly identify the variability as less for the sample size of 16. This was an indication that these students had a base knowledge of the sampling distribution and its characteristics.

To begin the second part of the task-based interview, students viewed the Random Rectangle simulation in *Fathom*, shown in Figure 6. The population of rectangles, labelled with their corresponding areas, is on the left and the graph of the areas of the total population of rectangles is in the upper middle. The Sample of Rectangles graph below that in the center displays a single random sample of rectangles. The Measures from Samples of Rectangles graph in the lower right displays the sampling distribution of the mean areas. Students were able to watch a demonstration that animated how each sample was taken from the population, graphed, and then the mean area from each sample was added to build the sampling distribution.

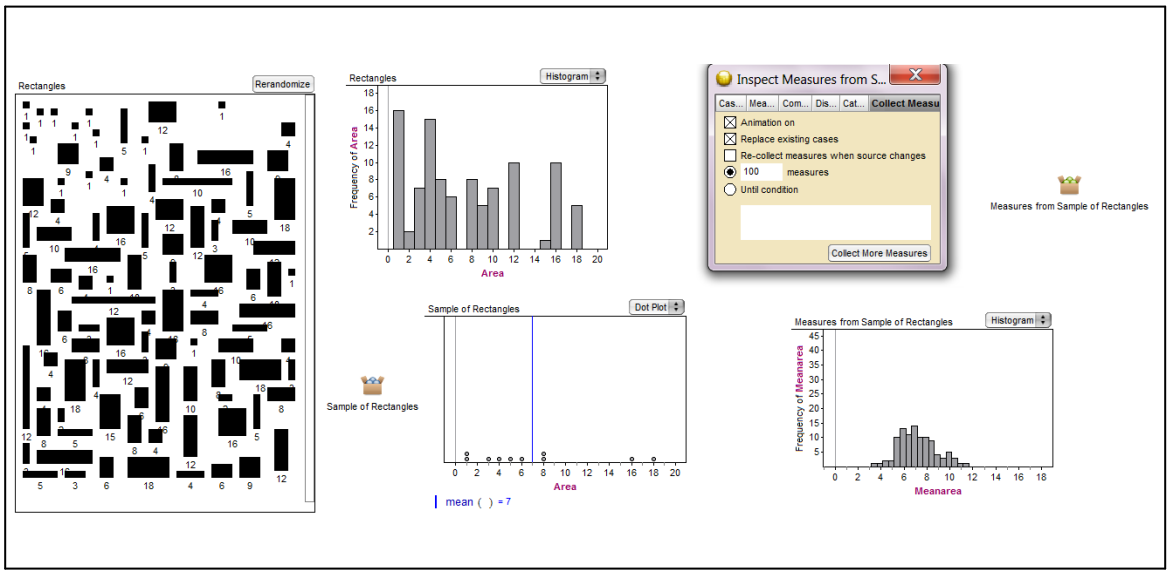


Figure 6. Screen shot of Random Rectangle simulation using *Fathom*

Following this demonstration, the students were shown three sampling distributions generated from this simulation for 100 samples of sizes five, 10, and 25 rectangles. When asked about how all three distributions compared to one another, six of the pairs referred to these distributions as

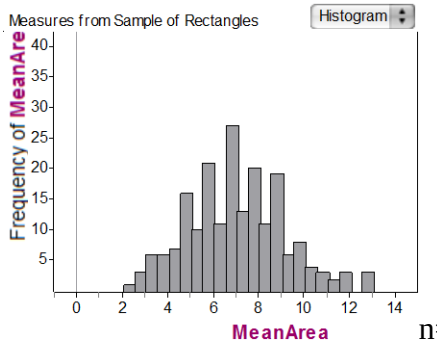
becoming more centered or having the same mean with five of them also referring to the decrease in variability as the sample size increased, as did Jared.

Interviewer: So we went from a sample size of 5, then to 10, now to 25. So how about this one [of sample size 25]?

Jared: This one's even more compact. The last one [of sample size 10] got all the way out to like 12. This one hasn't gone past 10 [referring to maximum mean area].

The remaining pair referred to the decrease in variability alone, mentioning the formula for standard deviation $\left(\frac{\sigma}{\sqrt{n}}\right)$ in support of this decrease. They recognized and commented on the changes in variability between the sampling distributions of different sample sizes.

Students were then asked what mean areas would be likely and which would be rare or unlikely for each of the sampling distributions. The students had no difficulty in identifying ranges of outcomes surrounding the peak of the approximately normal distributions as likely and those in the tails as rare as displayed in Figure 7. They demonstrated an understanding of the probabilities and variability associated with the normality of these sampling distributions of mean areas.



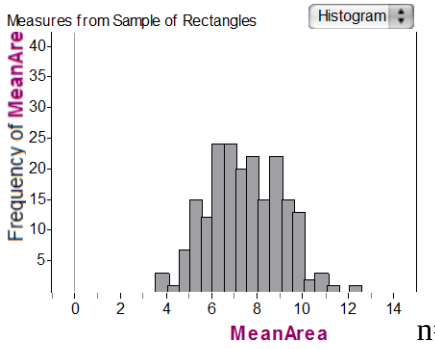
n=5

(1) Approximately what values of the sample mean for samples of size 5 would be reasonably likely?

3 - 10

(2) Rare events are defined as those that will occur less than 5% of the time. What values of the sample mean for samples of size 5 would you consider rare?

11 or higher and 3 or less



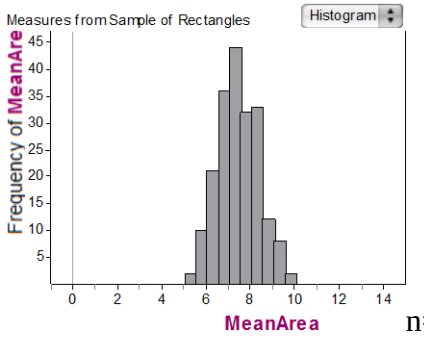
n=10

(1) Approximately what values of the sample mean for samples of size 10 would be reasonably likely?

5 - 10

(2) What values of the sample mean for samples of size 10 would you consider rare?

10 and higher and 4 and below



n=25

(1) Approximately what values of the sample mean for samples of size 25 would be reasonably likely?

6 - 9

(2) What values of the sample mean for samples of size 25 would you consider rare?

5 and less and 9 and higher

Figure 7. Example of student work in Part 2 of third task-based interview

In an effort to bring the previous concepts of normality and variability related to the sampling distribution together to make an informal inference, the interview concluded with the sampling distribution in Figure 8. In the second interview, students tossed small plastic houses to approximate the probability that a house would land upright when tossed. Students were shown this sampling distribution which was generated from 200 samples of 10 houses tossed, recording the proportion of houses landing upright. The interviewees were then asked if they could determine whether the probability that a hotel, slightly larger in size and more rectangular in shape, would land upright was the same as that for a house. Both the houses and hotels were available for students to manipulate.

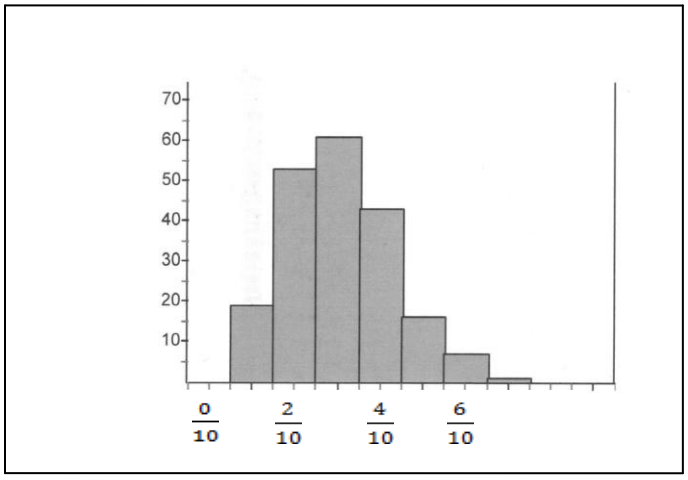


Figure 8. Sampling distribution of proportion of houses landing upright when tossed for 200 tosses of 10 houses.

Four of the seven pairs expressed that they would need to generate a sampling distribution exactly the same as the one they were shown for the houses. These pairs thought that they would need to toss 10 hotels 200 times to make a fair comparison, as expressed by Jared.

Interviewer: So is there a way that you could make some type of determination so that you could tell me that you think the probability [of the hotel] is the same [as the house] or that you think it's different.

Jared: Well, if we were to go through and roll them 200 times. I mean we could figure out what it would be compared to the houses.

Although these students had just demonstrated an understanding of the characteristics of the sampling distribution and had identified outcomes that could be considered likely and rare in relation to three sampling distributions, the majority of them believed that to accurately infer about the hotels, they would need to compare distributions and preferably, distributions of the same size. This was evident as another pair thought they would need to toss the hotels in the same manner as they had tossed the houses in the second task-based interview. At that time, they had tossed all 32 houses a total of five times. I interpreted this also as an indication that they wanted to compare measures of center as they were inclined to do when comparing distributions in the first task-based interview.

Since time constraints did not allow for replicating the sampling distribution, all of the pairs tossed 10 hotels: four of the pairs decided to toss 10 hotels 10 times, one pair tossed them three times, another tossed them twice, and the last pair tossed one hotel 10 times. It was not clear whether students were not grasping that the sampling distribution was a distribution of sample proportions so they could compare just one sample proportion to the distribution or whether they were confusing a sample proportion with a sample mean.

When students were drawing their conclusions about whether the probability that a hotel would land upright was the same as that for the houses, three of the pairs (two who tossed 10 hotels 10 times and the pair tossing one hotel 10 times) thought the probability would be the

same; however, one of those pairs was comparing their results of 0.17 to 0.19, their results from tossing the houses in the second task-based interview when they approximated that probability by sampling. The remaining four pairs stated that they thought the probability would be less even though their tosses resulted in percentages that were at or close to the peak of the sampling distribution. They were not taking the natural variability that could occur into consideration and, therefore, were not appropriately using their data as evidence in drawing their conclusions. A summary of the number of tosses of hotels by each pair and their concluding remarks about whether the probabilities for hotels and houses landing upright were the same are displayed in Table 24.

Table 24

Students' Concluding Remarks in Third Task-Based Interview

Pairs and Number of Tosses	Students' Concluding Remarks
April and Brian 10 (averaged)	April: Like we had more two's and it looks like this one has more two's. So I feel like it would have the same probability as the house. Brian: I'm still doubtful. What we found was about 25%. So about a fourth of the time it'll land upright, if not a little bit more than that. And for this [the sampling distribution] we have like 20%, less than 20%, so that's just me doing math in my head and I just don't think it's likely. Plus we only did 10 trials.
Caitlin and David 2	David: I think you'd have to try probably more times, many more times, but, as it looks right now it's about the same. Caitlin: Maybe a little bit less.
Emily and Fritz 1 hotel 10 times	Fritz: Yeah, it was pretty similar. Between 1 and 2 [houses landing upright out of 10] so I think that [their results] still validates that that's relatively the same.
Gabrielle and Jared 3	Jared: I'm figuring it's not going to be that far away. I think it's going to be roughly the same. It's maybe just a little bit less because it's weighted differently.

Laura and Mark 10 (averaged)	Laura: So we got 17%. I don't remember how many we did for the last one but that's like fairly close. And I think we did more or we like threw more houses last time. So I think that's pretty... I'd say they are about the same. Mark: Yeah, I'd say about the same.
Rachel and Steve 10	Steve: We didn't do nearly enough. I mean you did this 200 times, we did this 10 times so like you can't really say like, oh look what we did really quick and that refutes that. Rachel: Then I'd say it's different, but not by a lot.
Nathan and Pete 10 (averaged)	Pete: You gotta take more samples. …but with one trial, I think regardless of the outcome, you can't really compare that to what you got from this population [referring to sampling distribution]. It may fit into what you have seen. Like right here, this value right here, like 1, 2, [referring to peak in sampling distribution] ours was close so we could say yeah, it does compare similarly but I'm not going to bet my life on it.

The majority of the pairs could not make an accurate informal inference even though they had also identified likely and rare outcomes with the sampling distributions of mean areas of rectangles in the previous part of the interview. Additionally, the students had recently studied the normal distribution in their statistics classes which included the 68-95-99.7 rule of percentages of data within one, two, and three standard deviations of the mean. For the majority of them, this did not translate into the variability associated with this sampling distribution or that they could draw a conclusion based on a single sample of data.

Returning to the Makar and Rubin (2009) framework for thinking about informal statistical inference, all students were able to make an inference based on the data; however, the majority of them did not believe they had enough data to draw a conclusion about the probabilities. All students also used their data as evidence for making their inference with the majority of them comparing proportions rather than considering their results in relation to the sampling distribution presented to them. The probabilistic language used by the students was more evident in this task-based interview than in the previous interviews. They used phrases

such as “relatively the same”, “maybe a little bit less”, “it's different, but not by a lot”, or “I'm not going to bet my life on it.” The majority of the students expanded upon their probabilistic language.

Pre/posttests.

There were four questions on the pre/posttests involving drawing conclusions from a sample based on the corresponding sampling distribution. Questions 9 and 10 asked students to draw conclusions directly from a graph of the sampling distribution. For Questions 11 and 12, students were shown the population distribution and given the population mean. They were then asked to draw a conclusion based on the results of a random sample of size 50.

Questions 9 and 10 on the pre/posttest had students drawing informal inferences based on a sampling distribution for the proportion of heads expected when a fair coin was balanced on its edge 10 times. Marked on the sampling distribution in Figure 9 were the results of 0.7 heads and 0.9 heads from two different samples.

In Question 9, students were asked if it was reasonable to conclude that the coin was fair with a sample proportion of 0.7 heads. Eight of the students answered correctly on the pretest that this result, which was between 1 and 1.5 standard deviations from the mean, was reasonable. This improved to 10 students answering correctly on the posttest; however, one student changed a correct answer on the pretest to incorrect on the posttest.

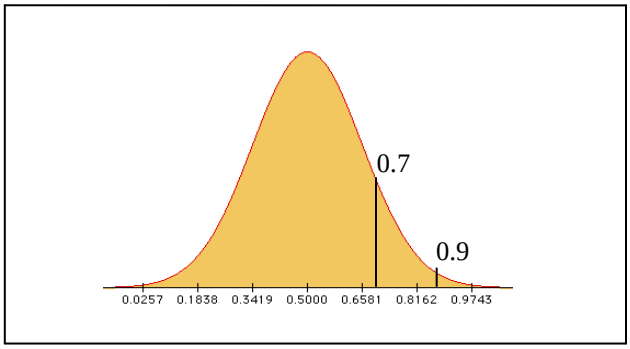


Figure 9. Sampling distribution for proportion of heads when fair coin balanced on its edge 10 times.

In Question 10, students were asked if it was reasonable to conclude that the coin was unfair with a sample proportion of 0.9 heads. Nine of the students answered correctly on the pretest and the posttest that this result, which was over 2 standard deviations from the mean, was reasonable. Two students changed their correct answers on the pretest to incorrect on the posttest. The results of Questions 9 and 10 are displayed in Table 25.

Table 25

Pretest/Posttest Results of Sampling Distribution Questions 9 and 10

Pretest/Posttest response	Number of students - Question 9	Number of students - Question 10
Correct/Correct	7	7
Incorrect/ Correct	3	2
Incorrect/Incorrect	3	3
Correct/Incorrect	1	2

Only six students answered both of these questions correctly and one student answered both questions incorrectly. Of the seven other students, three concluded that both results indicated that the coin was unfair while the other four students concluded that both results indicated that the coin was fair. These responses are consistent with reasoning students displayed when working on the hotel task. The majority of students were either incorrect or inconsistent in

their reasoning about the normality and variability associated with the sampling distribution. These two questions, with the sample results of 0.7 and 0.9 clearly marked on the sampling distribution graph, provided further evidence that most of the students were not able to appropriately use sample data as evidence because they were not fully considering the probabilities and variability associated with the sampling distribution.

For Questions 11 and 12 on the pre/posttests, the students were shown the left-skewed distribution of exam scores for a particular exam shown in Figure 10. The average exam score for this population was 74 out of 100 points. These questions added another layer of complexity in comparison to Questions 9 and 10 as the students had to reason about the sampling distribution given only the graph of the population distribution and its mean.

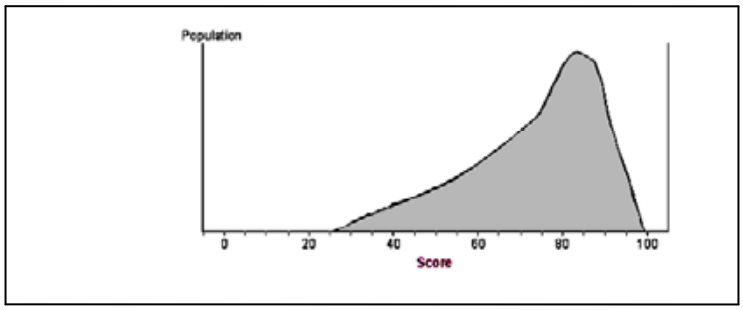


Figure 10. Population distribution of exam scores. Adapted from "A Framework to Support Research on Informal Inferential Reasoning," by A. Zieffler, J. Garfield, R. delMas, and C. Reading, 2008, *Statistics Education Research Journal, 7*(2), p. 51.

In Question 11, the students were asked if a current group of 50 students with an average of 78 points did better on average than expected for this exam. Based on the population graph given, students needed to approximate the standard deviation for the population and consider the related sampling distribution to determine whether this increase of four points was significant. Five students answered correctly on the pretest and seven answered correctly on the posttest that this four point increase did not indicate that the current group of 50 students in Question 11

scored better on average. However, three students changed their responses from correct on the pretest to incorrect on the posttest.

Question 12 asked the students if this higher sample average could just be due to chance. Nine students answered correctly that this higher score could be due to chance on the pretest and 10 answered correctly on the posttest. One student changed his answer from correct on the pretest to incorrect on the posttest. The results are displayed in Table 26.

Table 26

Pretest/Posttest Results of Sampling Distribution Questions 11 and 12

Pretest/Posttest response	Number of students - Question 11	Number of students - Question 12
Correct/Correct	2	8
Incorrect/ Correct	5	2
Incorrect/Incorrect	4	3
Correct/Incorrect	3	1

Taking Questions 11 and 12 together, six of the students responded correctly and were consistent in their reasoning by answering that this score could not be considered better than what could be expected and that the higher sample average score was due to chance. Three other students were consistent in their incorrect reasoning by answering that this higher sample average could be considered better than what could be expected and that it was not due to chance. The remaining five students displayed inconsistencies in their reasoning. Four of these five answered incorrectly that this higher sample average could be considered better than what could be expected but also that it was due to chance. The incorrect and inconsistent responses to these two questions is further evidence that the majority of these students were not drawing conclusions based on an understanding of the relationship between a population and the

corresponding sampling distribution including the probabilities and variability associated with its normality.

Overall the students had general knowledge of the sampling distribution. They knew it took on the shape of a normal distribution and they made references to the decrease in variability as the sample size increased. They also identified sample data values that would be considered likely and rare based on probabilities associated with the normality of sampling distributions. However, when it came to making a decision about the hotels in the last part of the interview, they did not completely reference what they knew. Instead, most of the pairs wanted to generate a sampling distribution for the hotels by tossing 10 hotels 200 times so it could be compared to the sampling distribution for the houses. Comparing measures of center was still the method of choice for these students in drawing a conclusion. The majority of them did not use the probabilities and variability associated with the normal distribution in drawing their conclusions nor did they show evidence of an understanding that the sampling distribution was a distribution of statistics to which they could compare a single statistic. The eight of fourteen students who answered either Question 9 or 10 incorrectly on the posttest provided further evidence of this. In Questions 11 and 12, there were eight students who either answered both incorrectly or who were inconsistent in their reasoning. For these two questions, the students needed to consider the effect of a sample size of 50 on the variability of the sampling distribution in relation to the population distribution. Although students had identified sampling distributions previously and discussed the effects of the sample size on variability, the majority of them did not use this information when asked to draw conclusions.

Procedural Knowledge in Formal Statistical Inference

As the student pairs worked through the final task-based interview on formal statistical inference, there was evidence that they were relying on their procedural knowledge rather than on a deep understanding of the implications of confidence intervals and hypothesis tests.

In Part 1 of the final task-based interview on formal statistical inference, students were shown 10 confidence intervals for the proportion of red beads from the same container they had used in the second task-based interview. In that interview they took samples and approximated the proportion of green beads in the container. These ten 90% confidence intervals were generated from 10 different samples taken from the container. Nine of the 10 intervals included proportions from approximately .35 to .65 red beads. One interval was different from the rest with a minimum of approximately .65 and a maximum of approximately .90 as the proportion of red beads in the container. I asked students if they could say anything about the proportion of red beads in the container based on these intervals. In Part 2 students were asked to take a sample of beads to find their own confidence interval. If they had difficulty with the formula, I provided them with it since the most important component was their interpretation of the interval. Students were then asked in Part 3 to conduct a hypothesis test with their sample to determine if they agreed with a student who believed that 70% of the beads in the container were red. Again, if students had difficulty with the formula for the test statistic, I gave that to them. I was analyzing their responses to determine if they could properly conduct the test but most importantly, whether they could interpret their procedure and results.

In the first part of the interview when students were asked to comment on the proportion of red beads in the container based upon the results of 10 confidence intervals, six of the seven

pairs concluded that the proportion of red beads was between .40 and .60 based on the intervals. This response from Caitlin and David was representative of the student responses.

David: A lot of them contain .4 and .5.

Caitlin: Some contain .6.

David: The average would probably be around, somewhere between these three obviously.

Caitlin: 40 to 60 %

David: Yeah.

Interviewer: What about the confidence intervals is helping you with that?

David: Most of them contain .4 and .5, and then most of them also contain .6. So it has to be in there, most likely.

Caitlin: Most of them.

The remaining pair stated that they thought the proportion was between .40 and .50 based on their approximation of the green beads from the second task-based interview and their visual analysis of the container of beads. I asked if the 10 confidence intervals would help them.

Interviewer: Do the intervals here help you with that or give you any insight into that [approximation of between .40 and .50]?

Rachel: Um. [pause] Yeah, I guess you could say that because of the fact that it's like, I didn't actually look at this [the 10 confidence intervals] when I like guessed it. Um, but it does make sense obviously like it, it looks like most of it is like .5 except for this one over here. [Points to one interval that does not contain .50] But like the ranges do seem to like correspond to my guess, around the area of half, if not less.

When this pair was asked if these intervals gave them any insight, one of the students noted that all but one of the intervals included .50 as a proportion of red beads. Therefore, they were confident with their conclusion.

Only two of the student pairs mentioned the one interval that was unlike the remaining nine intervals. April and Brian talked about this interval when they approximated the proportion of red beads to be between .40 and .60.

April: Like 40 %, except for this one. [Referring to the interval from approximately .65 to approximately .90]

Brian: Yeah that one's probably the only one that's off. Yeah, about like 40 to maybe 60%.

Interviewer: 40 to 60? Ok and again what about those intervals is helping you decide upon that?

April: Most of them include 4 to 6.

Brian: Of course except for like this one up here. [Referring to the interval from approximately .65 to approximately .90] Not sure what happened there but, you know, 90% confidence, the actual number's in there, between that range somewhere.

Brian indicates that there is something wrong with the one interval that does not contain the 40 to 60% of red beads like the other intervals. This is similar to Jared's comments about that interval.

Jared: Well a lot of people have it that it's right around 50%. This one here had it right around 80%. I don't know, seems like, I don't know if they knew what they were doing compared to everybody else.

Both Jared and Brian revealed that they thought there was a mistake made in formulating that interval. This was an indication that they were not realizing the variability that could occur when generating a confidence interval. Brian mentioned the 90% in his remarks but it was not clear that he was interpreting this correctly since he indicated that something was wrong with the one unusual interval. Whether the other five pairs of students were ignoring this interval because it was different or accepting of the variability among the intervals is unclear.

In forming their own confidence intervals in the second part of this task, two pairs took 10 samples to find the sample proportion, *p*-hat, for their interval, and one took three samples. These students were averaging to obtain their value for the sample proportion to use as their estimate for the proportion of red beads in the container. This was similar to the majority of the students' inclinations when concluding about the hotels in the third task-based interview. They wanted to take more than one sample to obtain their sample proportion, not believing that one sample proportion would provide accuracy. This could be seen when Caitlin and David began to work on their confidence interval by taking a sample of 32 beads with the paddle.

David: Thirty-two in here so, I got 20.

Caitlin: Yep. How many samples?

David: We'll take like 4 or 5.

Caitlin: 13 [red beads in the next sample of 32].

David: Yeah, 13.

Interviewer: Ok, can I ask you this? So what goes into forming a confidence interval? What are the pieces of the confidence interval?

Caitlin: Know if it's a z or t score.

David: You have to know it's standard deviation...

Caitlin: Standard deviation.

David: The standard deviation, sample and population.

Caitlin: Population or sample.

David: This would be sample.

Interviewer: Now here's the other thing to think about. This is, you're right, it's a sample, but it is also a proportion.

David: Yes it is.

Interviewer: So how many samples will you need to construct a confidence interval?

David: You should have a large number, [pause] but actually you wouldn't need many once you get your original proportion.

Caitlin: You could use just one.

David: You could use one.

Caitlin: That's all you really need.

When pressed to think about the components of a confidence interval, Caitlin and David realized they needed just one sample proportion. April and Brian, however, could not be convinced.

Brian: It's got to be a *p*-hat.

April: Yeah.

Brian: Eight, maybe 8 times doing this will be close to 2000 [total number of beads in the container], no it wouldn't. Eighty times doing this would be close to 2000 beads. But I don't want to do this 80 times. Um.

Interviewer: Yeah, how many times do you need to do it [take samples for the sample proportion]?

Brian: Well the more we do it [take sample proportions] the better it is [the sample proportion for the estimate in the confidence interval].

Interviewer: Alright, well think about these intervals [the 10 intervals from Part 1] that people created. How did they...

April: They did it 10 times.

Interviewer: These actually were 10 different groups in my class. So let's think about what is a confidence interval. How do you formulate the confidence interval?

Brian: Well we need the exp... well no we don't need expected… we need n.

April: x and n.

Brian: Yeah we're going to need x and we need mu. We're gonna need the proportion, not proportion...

Interviewer: Well think of it this way. So the confidence interval, what does it give you? What do you know from, well what you just did [in Part 1]...

April: A range?

Interviewer: Yeah you gave me a range right? So what you just did, you said well looking at these, it's somewhere between 40 and 60 [percent], ok. So if you're just looking for a range, how many samples do you need to take?

April: We could do 10.

Brian: Yeah I think 10 would be best, easy round number.

Interviewer: Well you could but you don't even need to do 10.

Brian: But it's more accurate if you do 10.

Interviewer: Well it is but remember that's the whole idea about that confidence interval. You're kind of getting that standard, that error, right, plus or

minus either way, right? So if you take one good sample and then give that margin of error, that's good enough for the confidence interval.

Brian: Well one good sample, it's, it could just be like a really lucky sample.

Interviewer: You're right.

Brian: Well you want to go under 10?

Despite the interviewer's efforts to remind them that the confidence interval was an interval estimate based on one sample proportion, Brian and April took 10 sample proportions of red beads and averaged them to obtain the sample proportion for their interval. Brian alluded to the Law of Large Numbers when he said that their sample proportion for the estimate in their confidence interval would be better with more trials. He also seemed to be confusing sample proportion and sample mean when he mentioned that they would need mu for their interval estimate. This was an indication that Brian was trying to minimize the variability in their estimate for the sample proportion. This pair was either not aware of the power of the sampling distribution used to formulate the confidence interval or unwilling to rely on that power.

After formulating their confidence intervals, five of the pairs wrote their results as they had been taught with a statement similar to, "We are 90% confident that the true proportion of red beads in the container is between [lower limit] and [upper limit]." The remaining two pairs verbalized this. The response from Laura and Mark was typical when asked what they had just found.

Laura: So it's between .4182 and .7068.

Interviewer: So can you tell me what that means? What's that tell you?

Laura: That, um, 90% of, or we can be 90% confident that...

Mark: It's between those 2 numbers.

Laura: Yeah, that the proportion of red beads is, the true proportion of the red beads is between those.

When asked what they meant by 90% confident, three of the pairs said there was a 90% chance the true proportion was in the interval. Caitlin and David's response was different.

Interviewer: So what does the 90% confident part of that mean?

Caitlin: It's kinda like...

David: Isn't it, I think, 90% of trials. That's what I thought.

Interviewer: Ninety percent of trials?

David: Would end up with a value in it.

Caitlin: Most of the trials.

Interviewer: So say that again. 90% of the trials would end up...

David: Would end up with a value in there.

Interviewer: In that interval?

David: Yes.

Caitlin: Or over multiple trials they found that 90% of the time your average is in that interval.

David indicated that he understood that the 90% referred to 90% of trials rather than a 90% chance and Caitlin referred to multiple trials; however, neither of them makes the distinction that over multiple trials of constructing the interval, 90% would contain the true population proportion. One student from yet another pair stated that the 90% meant that 90% of trials would result in this interval.

Interviewer: Ok, so this is a 90% confidence interval. What does the 90% mean?

Fritz: It means that we're 90% confident that the proportion of red beads drawn from a sample of 2000 beads is gonna be between 38.6% and 67.6%.

Interviewer: So what does that mean that you're 90% confident? For somebody who doesn't know statistics, what would you say?

Fritz: Like 90% of the time this would occur.

Interviewer: Ok describe that a little bit more. 90% of...

Fritz: Of trials like that you'd do. This would be the percentage that would result.

Fritz's interpretation was that 90% of the intervals constructed would be the same as the interval he and his partner calculated. The remaining two pairs replied that they did not know what being 90% confident meant.

Although all seven pairs of students provided a proper written or verbal concluding statement about their confidence intervals, the exact interpretation of the 90% associated with the confidence intervals was not clear for them. Three of the pairs interpreted the 90% as the chance the true population proportion of red beads would be in the interval; one pair interpreted the 90% to be the percentage of intervals constructed that would result in the exact same interval; two of the pairs stated that they did not know how to interpret the 90%; and the remaining pair was the only one to allude to 90% of multiple trials, but could not provide a complete explanation. This was further indication that the majority of the interviewees were not relating the 90% back to the role of the sampling distribution in formulating these intervals. The interviewees could properly calculate the upper and lower bounds of their confidence intervals but were not demonstrating an understanding that this 90% was the middle 90% of the area related to the normality of the sampling distribution.

In the final part of the last task-based interview, students were to use their sample proportion from Part 2 to test whether they agreed with a student who believed that 70% of the beads in the container were red. All seven pairs knew the procedure for conducting a hypothesis

test: calculate the test statistic, find the corresponding *p*-value, and then compare this *p*-value to the significance level. Only the three pairs of students from Rosemont drew the sampling distribution for their hypothesis test; however, two of the pairs could not identify what this distribution signified. Pete gave his interpretation.

Interviewer: Alright, let me ask you a question. What is that exactly that you drew right there [referring to the bell-shaped curve Pete had drawn]?

Pete: Well we said that, something like, our *p*-hat was .5. They're saying it's .7 so in comparison in our like bell curve, .7 is to the right 'cause it's greater than. We're looking to see if .7 is statistically significant. If it's that far away from the mean, then it's gonna be. So we find the *z*-score that's gonna give us this and then we find the area to the right of it and match up against the .05.

Interviewer: So what you have here, so what does that distribution represent? What is that?

Pete: Like a normal distribution, I don't know.

Interviewer: Normal distribution of what?

Pete: Ummmm, either the pop…, oh it's a proportion. So the normal distribution of the proportion if that's possible, I think.

Pete indicated that he knew how to use this normal distribution in the steps for completing the hypothesis test. However, Pete was unsure of what the normal distribution he drew represented and his partner did not offer an opinion on this. He was also confusing the *p*-hat of .5 and the claim of .7 (70% of the beads in the container were red) when he states that they were determining if .7 was statistically significant.

Four of the seven student pairs also had difficulty defining the *z*-value. These four pairs knew they were to use it to find a probability but did not relate the *z*-value to the normal distribution. The response from Steve and Rachel is representative of student responses.

Interviewer: So what's that -1.7 that you came up with there?

Steve: The *z*-score.

Interviewer: And what's the *z*-score? What's that tell you?

Rachel: Um, it tells you, um, I'm really not sure.

Steve: Yeah. We've been just like kind of, oh, *z*-score.

Rachel: Yeah but then it gives you the probabilities.

Interviewer: Right.

Rachel: So I don't know what, I'm not… The *z*-score gives you the probabilities, I think is the answer.

Similar to Steve and Rachel, the other three pairs knew they were finding a *z*-value for the test statistic but could not define it. This, again, indicated that they were not using the normality of the sampling distribution. If the *z*-value held little or no meaning for these students, it was likely the importance of the sampling distribution was an additional unknown abstraction for them.

After calculating the test statistic or *z*-value, all four of the pairs from Deerfield used their calculators to find the *p*-value rather than using tables. However, none of those pairs could define what the *p*-value represented.

Interviewer: So what you came up with is your *p*-value. And what is the *p*-value?

Fritz: Uh, how sure we are; like after this we'd write our concluding statement about it.

Fritz refers to the procedure used once the *p*-value is found. Gabrielle and Jared are challenged when asked about the *p*-value.

Interviewer: And then how about the *p*-value? What is the *p*-value? I know you're using it to compare to alpha, but what is it?

Jared: Um, ncdf is the probability.

Gabrielle: Proportion? No.

Jared: It's the probability. I've never really thought about how it connects the probability. It's like the probability that it's true or something along those lines.

Interviewer: You're right. It's a probability.

Jared: I know it's a probability. I just don't know how it relates really.

Gabrielle: I'm not sure either.

Jared: I've never thought of that.

Jared understands that the *p*-value is a probability, but he is unsure of what this is the probability. Similar to the *z*-value, the *p*-value did not hold meaning for these Deerfield students and was not acknowledged as an area under the sampling distribution.

All seven pairs were able to determine correctly, based on their sample proportion, whether they agreed with the student who believed that 70% of the beads were red. Two of the student pairs from Rosemont also returned to their confidence interval to confirm their results.

Interviewer: Ok, so then, just to answer that last question. Would you agree with this student [who believes the proportion of red beads in the container is 70%]?

Pete: No, I think it's a little far off what we assumed and what all those confidence intervals... there was only one that I saw that...

Nathan: Had the 7.

Pete: That's the only one that really agreed. None of them [the confidence intervals] agreed so I would not agree with this kid. Sorry.

Recognition of the relationship between the hypothesis tests they were conducting and the series of 10 confidence intervals for the proportion of red beads in the container provided insight that these two student pairs had moved further along the continuum of formal statistical inference than the other student pairs.

Throughout the final task-based interview on formal statistical inference, the pairs demonstrated their procedural knowledge in constructing confidence intervals and conducting hypothesis tests. However, when asked about the meaning of important components such as the percentage in the confidence interval, the bell-shaped curve drawn for the hypothesis test, or the *p*-value, they did not demonstrate an understanding that these centered around the sampling distribution and its normality.

Pre/posttests.

The first three formal statistical inference questions on the posttest involved interpreting a confidence interval. Students' responses demonstrated the difficulties they have with interpreting confidence intervals, particularly in identifying incorrect interpretations. The confidence interval estimated the average number of chocolate chips in a generic brand of chocolate chip cookies by stating that a 95% confidence interval for the average number of chips per cookie (18.6 to 21.3) was generated from a random sample. Of the three questions, Questions 13, 14, and 15, only Question 15 provided the correct interpretation of the confidence interval. In Question 15, the

interval interpretation was given as, "We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie." Thirteen of the interviewees answered correctly that this interpretation was valid. In Question 13, the interval interpretation was given as, "We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips." Eight of the interviewees answered correctly that this interpretation was invalid. In Question 14, the interval interpretation was given as, "We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips." Seven of the interviewees answered correctly that this interpretation was invalid. The students' responses, grouped by their responses to all three statements, are shown in Table 27.

Table 27

Interviewees' Responses to Confidence Interval Questions13 - 15 on Posttest

Number of Students Responding	Invalid Statement: We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips.	Invalid Statement: We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips.	Valid Statement: We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie.
4	invalid	invalid	valid
4	invalid	valid	valid
3	valid	invalid	valid
2	valid	valid	valid
1	valid	valid	invalid

While 13 of the 14 interviewees indentified the correct interpretation of the confidence interval, only four of the interviewees, all from Rosemont High School, answered correctly that the two other interpretations were invalid. Another four of these 13 interviewees, all from Deerfield High School, identified 95% of all possible sample means to be in this interval to be a valid interpretation as well. Three of the interviewees responded that the statement 95% of cookies would have between 18.6 and 21.3 chocolate chips was a valid interpretation in addition

to identifying the correct interpretation. Of the remaining three students, two responded that all three interpretations were valid and one student responded incorrectly to all three interpretations. Thirteen of the 14 interviewees responded correctly to the valid interpretation; however, nine of those 13 also responded that a second interpretation was valid as well. The valid statement in Question 15 represented the typical concluding statement given by all of the student pairs during the task-based interviews. With nine of the 13 students also identifying an invalid statement as valid, it was possible that the valid statement response had become a memorized concluding statement to a confidence interval with little true meaning for them.

For the next confidence interval question, Question 16, students needed to understand the relationship between sample size and the size of the confidence interval. They were asked which would produce a more precise or smaller interval, a sample of size 25 or a sample size of 64. Eleven of the interviewees answered correctly that the larger sample size would produce a more precise interval. Two of the interviewees responded incorrectly that the sample size of 25 would provide a more precise interval and one responded incorrectly that the intervals would have the same precision. This related to their understanding of the effect of sample size on the variability in the sampling distribution. As demonstrated in the interviews, students could discuss the effect of sample size on the variability of the sampling distribution but then could not always use that information.

Question 17 asked students to choose from four possible results when 110 statistics students each construct a 90% confidence interval. This question moved beyond choosing a concluding statement for a confidence interval as in Questions 13, 14, and 15. Six of the students correctly responded that this meant that about 90% of the 110 confidence intervals generated with random samples of the same size by each student in a statistics class would contain μ, the

true population mean. Four of the students incorrectly answered that about 10% of the raw scores in the samples would not be found in these confidence intervals. Of those four students, three correctly responded to all of the other confidence interval questions, Questions 13 through 16. Three of the students incorrectly answered that about 10% of the sample means would not be included in the confidence intervals. All of these three also answered the sample size question correctly in Question 16 but responded that more than one interpretation of the confidence interval in Questions 13 through 15 were valid. These results pointed to the inconsistencies in the majority of the students' understandings of confidence intervals and to their lack of depth of understanding. This provided additional evidence that for the majority of these students who responded correctly to Question 15 that the typical concluding statement was valid, they did not indicate that they knew what this meant when multiple confidence intervals were constructed with samples from the same population.

Students' responses to these five confidence interval questions indicated the inconsistencies in their understanding of confidence intervals, particularly in the correct interpretation of the percentage related to a confidence interval. This was consistent with their responses during the task-based interviews when they were not demonstrating an understanding that the 90% was the middle 90% of the area related to the normality of the sampling distribution. Together, the students' work during the task-based interviews and their responses to the confidence interval questions on the posttest provided evidence that they were able to construct these intervals and give an accurate concluding statement; however, they lacked the ability to properly interpret these intervals. I interpreted this to mean, for the majority of these students, their knowledge of confidence intervals was procedural.

The next two formal statistical inference questions on the posttest each gave an interpretation of the *p*-value for a new drug used to decrease vision loss. The students were to respond that the statement was valid or invalid. In Question18, eight of the fourteen interviewees correctly answered that the *p*-value was the probability of getting results as extreme as or more extreme than the ones in this study if the null hypothesis was true. Of the six responding incorrectly, three of them chose unsure as their answer. In Question 19, nine of the interviewees correctly answered that the *p*-value was not the probability that the alternative hypothesis was true. Only one of the students chose unsure as their response. Eight of the students responded correctly to both questions and five responded incorrectly to both questions. Although these students could not define the *p*-value during the task-based interviews, the majority of them could recognize the appropriate definition of the *p*-value in a written statement.

Question 20 was similar in that it asked for students to choose the most accurate interpretation of the *p*-value rather than whether a particular statement was valid or invalid. Six students chose the correct interpretation of the *p*-value. Seven of the eight students responding incorrectly chose the statement that the *p*-value was the probability that the null hypothesis was true. Taken together with Questions 18 and 19, five students answered all three questions correctly and four students answered all three incorrectly. Three of the students answered both questions 18 and 19 correctly, identifying a valid and an invalid statement for the *p*-value in the context of the results of a new drug test; however, did not choose the accurate general interpretation for the *p*-value from four possibilities in Question 20. The incorrect and inconsistent responses to these *p*-value interpretation questions provided evidence that for the majority of these students, their knowledge of hypothesis testing was primarily procedural.

In Question 21, students were asked to draw a conclusion based on the results of a hypothesis test. They were given the average length of fish in a lake, the population standard deviation, and the sample mean from a sample of 100 fish. Drawing the correct conclusion would require students to find the standard error of the sampling distribution and to consider how many standard errors were between the population mean and the sample mean. One student correctly answered this question. Ten of the students used the population standard deviation instead which resulted in the incorrect conclusion. This was an indication that these students were not drawing on the relationship between the characteristics of the sampling distribution, in particular the standard error, and its role in hypothesis testing. The majority of the interviewees were comparing a sample mean to the population distribution rather than to the distribution of sample means. The students were not explicitly asked to conduct a hypothesis test to answer this question; however, they had enough information to do so. There was no written work on the posttest papers that indicated the students conducted hypothesis tests. Without actually working through a hypothesis test, the majority of these students drew an inappropriate conclusion. This may be another indication that their knowledge of formal inference is primarily procedural. If not, it is quite possible that more of them would have used the standard error rather than the population standard deviation in drawing their conclusion.

The last question on the posttest, Question 22, asked students to draw a conclusion based on a *p*-value of .0025. Nine of the students correctly responded that this *p*-value provided strong evidence for the alternative hypothesis. Four of the students incorrectly responded that this *p*-value meant there was a small chance the alternative hypothesis was true in this context of comparing test scores to the national average. However, of these four students, three of them responded correctly that the *p*-value was not the probability that the alternative hypothesis was

true in the context of testing a new drug in Question 19, just three questions prior. These three students demonstrated inconsistencies in their interpretation of the *p*-value. It was not certain whether the nine students correctly responding to this question were thinking of this in terms of a small probability in the tail of the sampling distribution or whether this too came from a procedural knowledge of hypothesis testing.

Students' responses to the five hypothesis test questions revealed consistencies with their work during the final part of the task-based interview. Their difficulties in defining *z*-values and *p*-values in the interviews and inconsistencies in interpreting *p*-values on the posttest provide evidence that their knowledge of hypothesis testing is primarily procedural.

Differences by high school.

The quantitative analysis revealed that high school attended had an impact on the relationship between students' informal and formal inferential reasoning. In conducting the qualitative analysis, there were two instances in which the differences between students' responses differed by the high school they attended. The first occurred in the third classroom activity on sampling distributions and the second occurred during the last task-based interview on formal statistical inference.

In recapping the classroom activity on sampling distributions, three of the six interviewees from Rosemont who took part in the activity discussed drawing the sampling distribution. They each properly calculated the standard error and used that in drawing a conclusion about the claim in the two situations they were given. Of the seven interviewees from Deerfield who took part in the classroom activity, three of them who had the same teacher, drew the sampling distribution. Two of these students discussed finding *z*-values for the sample data and one approximated the number of standard deviations to draw a conclusion. The remaining

three students from Deerfield, all of whom had the other statistics teacher, constructed intervals without drawing the sampling distribution. The intervals corresponded to 1, 2, and 3 standard errors from the mean of the sampling distribution. They all used the 95% interval to determine if the sample data were within two standard deviations of the mean. Finding *z*-values and using formulas to construct confidence intervals indicated that some of the Deerfield students were using procedures of formal statistical inference to draw a conclusion for an informal statistical inference problem. The students from Rosemont did not use formal procedures or calculations indicating that they may have had more experience with drawing conclusions without the formalities of calculating *z*-values or constructing confidence intervals. It may be possible that they had more time to explore the relationship between the sampling distribution and what the probabilities associated with it indicated when drawing informal conclusions.

The second occurrence of differences between students from the two high schools took place when students were completing the hypothesis testing task during the last task-based interview on formal statistical inference. The three pairs of students from Rosemont drew the normal curve when completing the task. They did have some difficulties with defining what this normal distribution represented. Two of the pairs also referred to the corresponding confidence interval they had just found to confirm their results of the hypothesis test. These students had an understanding of the relationship between their 90% confidence interval and the results of their hypothesis test at the 5% level of significance. However, none of the four pairs of interviewees from Deerfield drew a normal curve as they completed the hypothesis test. In addition, they all found the *p*-value with their calculators. This may indicate that the Rosemont students had more graphical exposure to the role of the sampling distribution and its probabilities in formal

statistical inference. These two occurrences together may explain the significant correlation between informal and formal statistical inference found for the Rosemont students.

Summary

To summarize these results, I returned to my research questions for this study which were:

1) For students enrolled in an introductory statistics class:

a) Does their informal inferential reasoning develop?

b) If their informal inferential reasoning develops, what are the characteristics of this informal inferential reasoning as it develops?

2) What is the relationship between students' informal inferential reasoning and their formal inferential reasoning?

Development of students' informal inferential reasoning.

Using the three components of Makar and Rubin's (2009) definition as evidence of students' informal inferential reasoning, the majority of students' responses revealed that they: (1) made inferences based on the data, (2) used the data as evidence for their inferences, and (3) used probabilistic language to indicate a level of certainty in their inferences. In accordance with this definition, students' informal inferential reasoning did develop. However, how their inferences were based on their interpretations of the data gave further insight into the development of their informal inferential reasoning.

The majority of the interviewees compared only measures of center in the first task-based interview and this persisted through to the third task-based interview when they wanted to compare similar sampling distributions. The difficulties students had with interpreting variability in the first task-based interview also persisted through to the third task-based interview; however,

these difficulties manifested themselves differently. In the first interviews, when students were comparing distributions of test data, the majority of them referred to the variability in a variety of ways but then did not use it when formulating their informal inferences. During the third interviews when students were presented with an approximately normal sampling distribution, they did not use the variability and corresponding probabilities associated with any normal distribution when formulating their informal inferences. It is possible that without a true sense of how variability can differentiate two distributions, students were not willing to rely on the variability and corresponding probabilities of the normal distribution.

Relationship between students' informal and formal inferential reasoning.

I refer back to my particular findings to discuss the relationship between students informal and formal inferential reasoning. As previously stated, the difficulties with interpreting variability persisted for the majority of students. This translated into uncertainties in drawing conclusions based on the sampling distribution. Since formal inference is based entirely on the power of the sampling distribution, it was not surprising that students then could not fully use that power when constructing confidence intervals or conducting hypothesis tests. Instead, they primarily relied on their procedural knowledge for formal statistical inference. Further insight into the relationship between informal and formal statistical inference was gained by referring back to the quantitative analysis which revealed a difference in this relationship by high school attended. For the Deerfield students, using formal procedures to find *z*-values and construct intervals to draw informal inferences with the sampling distribution may have contributed to the lack of a relationship between their informal and formal inferential reasoning on the posttests.

In the final chapter, I will discuss the implications of these results together with the results of the quantitative analysis. I will also discuss the limitations of this research and possible areas of further research.

Chapter 6 - Discussion and Conclusions

In this study I investigated introductory statistics students' informal inferential reasoning in two ways. First, I attempted to determine if students' informal inferential reasoning developed over the course of their introductory statistics class; and, if so, what were the characteristics of that informal inferential reasoning. Second, I explored the relationship between students' informal and formal inferential reasoning to determine what characteristics of students' informal inferential reasoning corresponded to their formal inferential reasoning. I did this with an analysis of a pre/posttest assessment of 136 introductory statistics students and a series of four task-based interviews with seven pairs of students.

To frame this study, I used a series of informal statistical inference questions and tasks drawing on the task framework proposed by Zieffler and colleagues (Zieffler et al., 2008) for the study of informal inferential reasoning. The three types of questions and tasks I chose were aligned with the curriculum of a typical introductory statistics course. These questions and tasks provided insight into students' informal inferential reasoning as it developed during their study of introductory statistics. These questions and tasks included: comparing distributions of data (drawing on the research of difficulties students encounter with the concepts of variation and distribution in descriptive statistics); sampling and estimating a probability to explore students' understandings of randomness, the law of large numbers, and probabilistic thinking; and inferring about a population based on a sample of data which stems from students' difficulties in understanding the sampling distribution. The questions on the pre/posttest and the sequence of tasks for the task-based interviews provided snapshots of students' growth throughout their study of introductory statistics and in three key topic areas linked to the development of their informal inferential reasoning.

In analyzing students' responses during the first three task-based interviews, I used the essential principles for informal statistical inference developed by Makar and Rubin (2009). Evidence of students' informal inferential reasoning was determined by the extent to which they (1) made inferences based on the data, (2) used the data as evidence for their inference, and (3) used probabilistic language to indicate a level of certainty in their inference. This framework provided a starting point for analyzing students' responses; however, I needed to delve much further into students' responses to find the other recurring themes. This framework did not fully account for students' correct or incorrect inferences. For example, there were instances when students exhibited all three principles of informal statistical inference but inferred incorrectly. Comparing students' reasoning for their correct and incorrect inferences provided the basis for most of the qualitative analysis.

I will next highlight the main quantitative and qualitative findings and then discuss how they provided answers to the research questions for this study. This will be followed by the limitations of this study. I will conclude with the implications of this research for practice and possible areas for further related research.

Quantitative Findings

The quantitative analysis of the pre/posttests administered to the 136 introductory statistics students revealed three main findings regarding students' inferential reasoning. The first finding involved the significant improvement in students' informal inferential reasoning from the pretest to the posttest. Results for students in each high school and for the strong informal inferential reasoners (those scoring above average on the pre and posttest informal statistical inference questions) showed significant improvement in their overall informal inferential reasoning. The second finding included the relationships that existed between students' informal

and formal inferential reasoning on the posttest. For the Rosemont students and the strong informal inferential reasoners, their overall informal inferential reasoning was significantly related to their formal inferential reasoning. The final finding involved the 14 interviewees who demonstrated stronger formal inferential reasoning than the remaining 122 students based on their posttest scores.

Significant improvements in informal inferential reasoning.

By comparing students' responses to the informal statistical inference (ISI) questions on both the pretest and the posttest, it was evident that students' informal inferential reasoning did develop due to their regular classroom instruction. Students from each high school and the strong informal inferential reasoners had significant gains in their overall informal statistical inference scores. The following are the results in each of the three subcategories of informal statistical inference.

The first subcategory of questions on the pre/posttests measured students' informal inferential reasoning when comparing two distributions of data for which neither high school showed significant improvement. To gain insight into why students did not show improvement in this subcategory of comparing distributions, these scores were compared to the scores in the other two subcategories on the pretest. This analysis revealed that both Deerfield and Rosemont students scored significantly higher in this subcategory of comparing distributions than in the other two subcategories of informal statistical inference questions on the pretest. This may have created a ceiling effect since there were only four questions in the subcategory which did not allow for significant improvement for either high school. As I stated in the quantitative analysis, it is likely that their previous mathematics instruction prepared them for comparing distributions.

The second subcategory of questions on the pre and posttests measured students' informal inferential reasoning when drawing conclusions based on sampling and estimating a probability. Students in each of the high schools exhibited statistically significant improvement in their reasoning about how sampling and probability are both used in drawing an informal inference. The strong informal inferential reasoners also showed significant improvement in this category. Knowledge of sampling and the probabilistic reasoning associated with it are required for formal inferential reasoning.

The third subcategory of questions on the pre and posttests measured students' informal inferential reasoning when comparing data from a single sample to the sampling distribution of all such samples. Deerfield students and the strong informal inferential reasoners exhibited statistically significant gains in their responses indicating that they improved in their ability to draw an informal inference by comparing a sample of data to its related sampling distribution. Knowledge of the sampling distribution and how it is used to draw conclusions provides the basis for formal statistical inference.

These subcategories offered insight into the characteristics of students' informal inferential reasoning as it developed. Students' previous instruction supported their inferential reasoning when comparing distributions; and all students' informal inferential reasoning developed when drawing conclusions based on sampling and estimating probabilities. When drawing informal inferences with the sampling distribution, Deerfield students showed development in this area as did the group of strong informal inferential reasoners.

Relationships between informal and formal inferential reasoning.

In examining the relationship between students 'overall informal and formal inferential reasoning, a significant correlation existed for the students at Rosemont High School ($r = .35$)

but not for the Deerfield High School students. The results from the task-based interviews shed light on possible reasons this difference between high schools occurred. For the strong informal inferential reasoners, there was also a significant positive correlation between their overall informal statistical inference scores and their formal statistical inference scores ($r = .36$). This was largely due to the significant positive correlation between their subscores for inferring about a population based on a sample of data and their formal statistical inference scores. This was an indication that students who began as and remained strong informal inferential reasoners, particularly those strong in inferring about a population based on a sample of data, were also strong formal inferential reasoners. In addition, these 46 students were also stronger formal inferential reasoners than the other students at the end of their introductory statistics course.

Interviewees' formal inferential reasoning.

At the completion of the study, the 14 interviewees scored significantly higher than the remaining 122 students on the formal statistical inference questions on the posttest. This may have been due to their extra work during the task-based interviews. The structure and the timing of these interviews gave the interviewees extended experiences in each of the three informal inferential reasoning subcategories of comparing distributions of data, sampling and estimating probabilities, and inferring about a population based on a sample of data. The qualitative findings of the task-based interviews provided more insight into the interviewees' informal and formal inferential reasoning.

Qualitative Findings

The qualitative analysis of the task-based interviews revealed three findings in regards to the research goals for this study. The first finding emerged during the analysis of the first task-based interviews completed by the student pairs. Given five different sets of two histograms of

test scores, the majority of the students relied on the mean and/or median to determine if the classes scored equally well or if one class scored better overall than the other. This occurred even though there were differences in the variability in the test scores for these classes. The second finding surfaced with the third set of task-based interviews when students were asked to draw informal inferences based on a sampling distribution. The majority of the interviewees were not comfortable using a single sample to draw a conclusion and the majority of them did not use the probabilities associated with the normality of the sampling distribution when comparing samples to the related sampling distribution. The last main finding emerged during the final task-based interviews on formal statistical inference. During the interviews, it became clear that the majority of the students were relying on their procedural knowledge in constructing a confidence interval and conducting a hypothesis test.

I will discuss each of these findings in light of the related literature and return to my research questions to highlight common threads that connected these findings.

Reliance on mean or median when comparing distributions.

During the first task-based interviews, the seven pairs of students compared distributions of class test scores. It became clear that they were recognizing the differences in variability among the five sets of distributions they were comparing; however, the majority of them were not including these differences in variability when deciding which class scored better. They primarily relied on the means or medians in drawing their informal inferences.

These results provide evidence that the majority of the interviewees, similar to the preservice teachers in Makar and Confrey's (2005) study, were aware of variation; however, this did not ensure that they grasped the significance of the variation. The interviewees used several phrases to describe the differences in variation between the pairs of distributions they were

comparing such as “more spread out” or “more dense”. However, most of the students relied on an equivalent mean or median to infer that the classes scored equally well. This could have been due to the timing of the first task-based interviews which took place early in their study of introductory statistics. It is possible that there had not yet been the opportunity for the teachers to have explored deeply the implications of differing variabilities between data sets. Additionally, the statistics instruction that students would have been exposed to prior to this introductory statistics class would have primarily focused on measures of center rather than variability. Therefore, recognizing variation and also grasping its significance by using it to draw conclusions would indicate a development in informal inferential reasoning when comparing distributions of data.

As the 14 interviewees compared distributions of data in the first task-based interview, there was evidence that they were responding at the second level of the unistructural-multistructural-relational (U-M-R) learning cycle used by Watson and Moritz (1999) in their study of informal inference with students in grades three through nine. At this second level, multistructural, students were seeing several characteristics in the data they were comparing (i.e. center, spread, skewness); however, these characteristics were not working together when making an inference. Watson and Moritz found that this often created a conflict for students that they were not able to reason through. For the 14 interviewees in my study, there was evidence that they recognized the characteristics of center, spread, and skewness in the distributions they were comparing, but the majority of the interviewees relied on the measure of center to draw inferences about those distributions. This did not cause a conflict for most of the interviewees; only three students remained undecided as to whether one class scored better or they scored equally well when presented with two symmetrical histograms with the same number of test

scores and the same mean and median but different variability. These three students knew that the differences in variability should be taken into consideration; but they were unclear as to exactly what this meant in a statistical sense when comparing the two distributions.

Referring to Wild and Pfannkuch's (1999) four-dimensional framework for statistical thinking, in their Dimension 2: Types of Thinking, variation is a fundamental component of statistical thinking. They described four characteristics in thinking statistically about variation, two of which pertain to the statistical thinking needed for the comparing distributions task: noticing and acknowledging variation; and explaining and dealing with variation. Evidence from the task-based interviews suggested that, except for three students, the remaining 11 interviewees only focused on one of these components, noticing and acknowledging variation. In the majority of instances, as stated previously, the interviewees pointed out the variation; however, they proceeded to base their inferences on the measure of center alone. Caitlin, Nathan, and Pete were further along the spectrum in their informal inferential reasoning when comparing distributions. They showed evidence of attempting to explain and deal with the differences in variation between the two distributions of scores. The conflict for Caitlin, Nathan, and Pete was whether the difference in variability was enough cause for determining that the class with less variation in their test scores had scored better. This may have been due to the small sample size of the class data of test scores. It is possible that a difference of two test scores was simply not convincing enough to conclude that the class with less variability scored better. It is also possible that the majority of the students understood scoring "better" as a question about the center. I will discuss this in more detail in the limitations section of this chapter. In addition, as stated previously, these first task-based interviews took place early in the introductory statistics course; therefore these students may have had little experience in their classrooms with ways of explaining and

dealing with variability. This is an area that deserves further research and I will discuss that later in this chapter as well.

In the fifth part of the interview when students needed to reason proportionally about the variability in making an inference, four students did not and inferred that the class with the largest set of data scored better overall even though it had the same mean and median with larger variability. Caitlin reverted to using just the measure of center as did six other students. These students simply did not take the variation into consideration. Only Nathan and Pete showed a progression in their informal inferential reasoning when they inferred correctly based on the measure of center and the variation. Nathan focused on the effect of low scores on each of the distributions. He stated that the Orange class would not be greatly affected by a low score of two, for example; however, the Grey class would be affected by a low score. This represented another way of expressing the variability in terms of proportional reasoning due to the different class sizes. His partner Pete agreed with this thinking and went on to mention that data with a smaller range would show a "very tight" boxplot. Nathan and Pete are demonstrating a more global view of the differences in variation that they are seeing than the other interviewees.

Drawing conclusions with the sampling distribution.

When it came to drawing informal inferences with the sampling distribution in the last part of the third task-based interviews, the majority of the interviewees exhibited uneasiness in relying on a single sample of data. Overall the students had a general knowledge of the sampling distribution. In most college level introductory statistics courses, the focus is contained to sampling distributions that are approximately normal. They knew it took on the shape of a normal distribution and they made references to the decrease in variability as the sample size increased. They also identified sample data values that would be considered likely and rare based

on probabilities associated with the normality of sampling distributions. Rather than rely on a single sample proportion, students wanted to compare distributions of samples proportions by generating another sampling distribution; however, this could not be done due to time constraints. Additionally, although their sample results were close to the peak of the sampling distribution, the majority of them did not use the probabilities and variability associated with the normality of the sampling distribution in drawing their conclusions. It is possible that since the interviewees already had experience with tossing houses, it seemed reasonable to toss the hotels in the same manner. That inclination may have been too strong to put aside and rely on the sampling distribution that had been given to them. The fact that they did not generate that sampling distribution themselves may not have led them to view that as a sampling distribution with all of the characteristics they had just reviewed in the previous parts of the task.

In light of the research on students' understandings of the sampling distribution, Saldanha and Thompson (2002) found the majority of their students compared a single sample statistic to the population parameter rather than to the sampling distribution of all such statistics when asked to determine if it was unusual. These authors used computer simulations in which students took many samples from a variety of populations with known parameters. When asked if they could determine if the probability that a hotel would land upright was the same as that for a house, the students in my study were also hesitant to compare their sample proportions to the sampling distribution presented to them. They were not able to compare to a population parameter as in Saldanha and Thompson's study since the true p was not known and there was no clear theoretical distribution for the population of houses landing upright when tossed. One of the pairs, however, correctly drew a conclusion by comparing their sample proportion of hotels landing upright to the proportion of houses landing upright that they approximated in the second

task. Saldanha and Thompson found that students were able to reason about the variability among samples; however, this reasoning did not necessarily translate to the variation that existed in sample statistics. The students in my study wanted to take many samples, exhibiting similar reasoning, to diminish the effects of the variability that could occur between samples. However, like the students in Saldanha and Thompson's study, this realization of variability did not translate to the variability associated with the sampling distribution of sample statistics. This prevented them from having confidence that they could draw a conclusion based on just one or even a small number of samples. Taking many samples with the hotels may have felt like a much more concrete method for inferring about the probability that a hotel would land upright. Possibly relying on this sampling distribution that they did not create and could not be sure of the circumstances under which it was created, was viewed as another dimension of variability that they chose to avoid.

Pratt and colleagues (Pratt et al., 2008) examined local and global thinking when students reasoned informally. Using software designed by Pratt for the study, students had the ability to add to an existing sample or generate a new sample. In either situation, students tended to focus on the changes in subsequent displays of the data, similar to the students in Saldanha and Thompson's study. At times when they did express a global understanding by referring to the stability found when considering all of the samples, they were still frustrated by the fluctuations they saw in the individual samples. This suggested that an important aspect of informal inference is in finding the invariance that is present even among all of the local changes. The majority of the students in my study viewed their samples in a local sense. They were concerned about those samples being accurate enough to draw a conclusion. They were not viewing their samples in a global sense when comparing them to the sampling distribution; and, therefore, could not rely on

the variability associated with its normality. If they had, a result of 1/10 hotels landing upright, for example, would not have triggered a response that this result differed from the 2/10 peak of the sampling distribution for the proportion of houses landing upright as it did for several of the interviewees.

Procedural knowledge in formal statistical inference.

The final finding emerged as the student pairs worked through the last task-based interviews on formal inference. It became clear that they were primarily depending upon a procedural knowledge while constructing a confidence interval and conducting a hypothesis test. In addition, to begin the tasks for the final interview, the majority of them did not want to rely on one sample for constructing a confidence interval. This was similar to their work when drawing an informal inference based on the sampling distribution. It was clear that students were still not relying on the power of the sampling distribution; therefore, they had to depend upon their procedural knowledge in completing the formal inference tasks.

This reliance on procedural knowledge for formal inference was evident in the interviewees' responses to the confidence interval questions on the posttest. Thirteen of the 14 interviewees responded correctly to the valid interpretation; however, nine of those 13 also responded that a second incorrect interpretation was valid as well. These results were consistent with the findings of delMas, Garfield, Ooms, and Chance (2007) as they were developing CAOS, an assessment to capture students' understandings of statistical concepts including their probabilistic reasoning at the completion of an introductory statistics course. They found that for both *p*-values and confidence intervals, students demonstrated that they could recognize a correct interpretation but also indicated that incorrect interpretations were valid. Additionally, they misinterpreted the *p*-value as the probability that a treatment was not effective and a confidence

level as the percentage of sample data falling in the interval. The interviewees for this study were questioned about *p*-values; and, although they could not define *p*-value during the task-based interviews, the majority of them could recognize a valid interpretation of the *p*-value on the posttest. However, when asked to choose the most accurate interpretation of the *p*-value in another posttest question, only six students chose the correct interpretation of the *p*-value. Seven of the eight students responding incorrectly chose the statement that the *p*-value was the probability that the null hypothesis was true.

This result is consistent with research conducted by Castro Sotos, Vanhoof, Noortgate, and Onghena (2009). They found students were prone to confusing the probabilistic meaning of the *p*-value. Twenty-one percent of their students identified the *p*-value as the probability of the null hypothesis and 16% identified it as the probability of incorrectly rejecting the null hypothesis. Castro Sotos and colleagues also found students showed a lack of understanding of exactly what the result of a hypothesis test means. Twenty percent of the students in their study responded that a hypothesis test proves or disproves the null hypothesis. In contrast, the students in my study did not demonstrate this as they all were able to draw an appropriate conclusion based on their hypothesis tests.

Conclusions

Through the quantitative analysis of the pre/posttest administered to all 136 introductory statistics students, there were indications that students' informal inferential reasoning did develop with their regular classroom instruction. This was seen in the significant improvement in student responses to the informal statistical inference questions particularly when sampling and estimating a probability and inferring about a population based on a sample of data.

Analysis of the task-based interviews revealed additional insight into the characteristics of the interviewees' informal inferential reasoning. During the first interview when students were comparing distributions of class test scores, they focused on the mean and/or median to draw their conclusions. Many of them also referred to the differences in variability between the distributions. During the third task-based interview when students were asked to draw a conclusion based on a sampling distribution, the majority of them wanted to return to comparing distributions by generating another sampling distribution for the proportion of hotels landing upright. I interpreted this as the comparing of means or medians providing these students with a higher degree of certainty than relying on the normality of the sampling distribution. This presented a barrier for them in using the sampling distribution to draw a conclusion.

During the second task-based interviews when students took their own samples to approximate population parameters for the proportion of houses landing upright when tossed and for the proportion of green beads in the container, they demonstrated a clear understanding of the law of large numbers. Students had difficulty in transferring this concept to the sampling distribution which was generated from a large number of samples. During the third task-based interviews, they held on to their accurate concept of the law of large numbers, wanting to take many more samples to compare to the sampling distribution. While this was an attempt to reduce variability, which is a correct intuition, using the known variability that exists in the sampling distribution of all such samples is a more efficient method. This same type of reasoning appeared during the final task-based interviews on formal statistical inference when many of these students wanted to take several samples for the proportion estimate to construct a confidence interval. These characteristics then provided insight for addressing the relationship between students' informal inferential reasoning and their formal inferential reasoning.

Whether students' difficulties in relying on the sampling distribution stemmed from not considering the sampling distribution a powerful tool generated from a large number of samples or whether they found more certainty in comparing measures of center, this prevented them from making the necessary connections for formal statistical inference. The sampling distribution is the key to formal statistical inference and without a deep understanding of its power, relying on it for formal statistical inference is not possible. Therefore, as students learned the procedures for formal statistical inference, the sampling distribution was likely not paramount. Those procedures for formal statistical inference and the underlying concepts related to the sampling distribution may have had little connection for students. Therefore, concepts like the *p*-value remain unclear. This sheds light on the reason that introductory statistics students experience difficulty interpreting *p*-values. Without this connection, they could only rely on their procedural knowledge.

The quantitative analysis revealed that high school attended had an impact on the strength of the relationship between students' informal and formal inferential reasoning. In conducting the qualitative analysis, there were two instances in which the differences between students' responses differed by the high school they attended. The first occurred in the third classroom activity on sampling distributions. Finding *z*-values and using formulas to construct confidence intervals indicated that the Deerfield students were using procedures of formal statistical inference to draw a conclusion for an informal statistical inference problem. The students from Rosemont did not use formal procedures or calculations indicating that they may have had more experience with drawing conclusions without the formalities of calculating *z*-values or constructing confidence intervals. The second occurrence of differences between students from the two high schools took place when students were completing the hypothesis testing task

during the last task-based interview on formal statistical inference. The three pairs of students from Rosemont drew the normal curve when completing the task. Two of these three pairs also referred to the corresponding confidence interval they had just found to confirm their results of the hypothesis test. These students had an understanding of the relationship between their 90% confidence interval and the results of their hypothesis test at the 5% level of significance. However, none of the four pairs of interviewees from Deerfield drew a normal curve as they completed the hypothesis test. In addition, they all found the *p*-value with their calculators. These two occurrences together may explain the significant correlation between informal and formal statistical inference found for the Rosemont students. This would suggest that waiting to introduce formal procedures would benefit students' informal inferential reasoning in support of their formal inferential reasoning.

Limitations of this Research

There were several limitations of this study which I will describe. The first was the nature of the high schools involved in the study. There were only two suburban high schools that took part, therefore the results may not generalize to a large number of high school introductory statistics classes with students taking the course for college credit. However, the study did reveal results that were consistent over the different classes of students, which strengthened the external validity of the study.

Another limitation of this study involved the selection of the participants for the student interviews, posing a threat to external validity. These students were not randomly selected, but instead were selected with the assistance of the classroom teachers. To ensure that these task-based interviews would be productive, I asked the teachers to recommend students who would be inclined to discuss freely their ideas as they worked through the tasks. In addition, I asked the

teachers to recommend students they believed were at differing levels of mathematical achievements as evidenced by previous accomplishments as well as accomplishments in the first several weeks of their introductory statistics class. These parameters were subjective; however, the students chosen by these teachers all freely took part in the four task-based interviews with these interviews proving to contain worthwhile conversation. These criteria for selecting the interviewees allowed for a comprehensive analysis of their understandings of key statistical concepts and the development of their informal inferential reasoning.

Having observed these statistics classes only three times (once during each of the classroom activities), I was limited in my knowledge of the nature of the instruction that took place on a daily basis in these classrooms. Since the scope of my study was focused on students' informal inferential reasoning, I chose to concentrate on students' work rather than observing the classroom teachers. Clearly the classroom instruction had an impact on the students' statistical reasoning and the addition of this data would have been beneficial. Time and resource constraints did not make the collection of this data feasible. While this posed a threat to internal, external, and statistical conclusion validity, the fact that these classes were introductory statistics classes offered consistency in the curriculum topics that provided support for the task-based interviews and quantitative assessments.

After analyzing students' responses, I would make some changes to the tasks used in the task-based interviews. I would first make a change to the five pairs of histograms the interviewees compared in the first task-based interview. I would increase the sample sizes of these distributions so the distributions would be more consistent with those displayed on the pre/posttest. I think that the small sample sizes may have made the differences in variability appear too minimal for many of the students to seriously consider. Another change that I would

consider is including comparing distribution questions similar to those on the pre/posttest. In particular, in Questions 3 and 4 on the pre/posttest, students were shown four sets of boxplots comparing running times of athletes split according to those who were in a weight training program and those who were not weight training. Students were asked to identify the set of boxplots providing the most convincing evidence and then the least convincing evidence that the weight training program was effective. This type of question requires students to compare medians and to compare the differences in variability. I think an interview with questions of this nature would provide more insight into how students are reasoning about differences in variability.

There is also a change that I would make to the bead task in the second task-based interview. During that probability task, as students sampled from the container of beads to determine the proportion of green beads, one pair was approximating this proportion initially by sight. The container was transparent and rather than taking samples, this pair examined the container to approximate the proportion of green beads. I would modify this container so that would not be a possibility. This would ensure that this task aligned with the tossing pigs classroom activity and the tossing Monopoly houses task in that same interview. In those instances, the only method for approximating the probabilities was to take samples as there was no clear theoretical probability.

One more change I would make is to the third task-based interview when students worked with the sampling distribution of houses. Rather than give them that distribution, I would ask them to generate it by taking samples. This would take more time to complete, but I think it may enhance students' understanding that this sampling distribution has the characteristics they know exist for sampling distributions in general.

Implications for Practice and Future Research

The statistics education community would benefit from studies that analyze teaching strategies and sequences to determine if students' informal inferential reasoning could develop to a level that would facilitate a seamless transition to formal statistical inference. This is the nature of the future research I suggest.

Based on the results of the first task-based interviews and the concept of variability, I believe students would benefit from conversations about when variability is an important factor in comparing distributions. For example, if students were asked during the first task of comparing class test scores which class they would choose to team up with for a math competition, this might draw out more of their reasoning about variability and its importance. Students could identify particular circumstances when consistency is important and others when it is not. This may also draw out the remaining two components of statistical thinking about variation proposed by Wild and Pfannkuch (1999), measuring and modeling variation for the purposes of prediction, explanation, or control and investigative strategies. A study designed to engage students in practical ways of dealing with variability may illuminate teaching strategies that will support their informal inferential reasoning.

It would be worthwhile to investigate how students would feel about the need for more samples in a task like the one with the Monopoly houses and hotels in the third task-based interview if they were able to generate another sampling distribution to compare as many of the pairs initially wanted. Allowing students to explore their notions with the sampling distribution and the sample size may help them to understand that this will not necessarily provide them with more certainty in their conclusions. Discussions about efficiency as well as accuracy in data collection could take place. These students likely do not have any practical experience with

working within budget or time constraints for using data to draw conclusions; therefore, recreating the sampling distribution to draw a comparison to make a decision may have seemed like the only certain method. Students need experiences that will help them to understand that one sample can be enough to draw an inference with the sampling distribution. Therefore, a study allowing them to explore their notions regarding samples and comparing them to the sampling distribution may assist the statistics education community in developing activities to support informal inferential reasoning that will, in turn, support their formal inferential reasoning.

In addition to giving students more experiences with efficiency in data collection and drawing conclusions, more time could be spent on inferring with the sampling distribution before heading into formal statistical inference. This may help students to see the need for the significance level in hypothesis testing as the "cut off" between reasonably likely and unlikely. This may also enhance their understanding of the significance of *z*-values and then *p*-values in formal statistical inference. Research efforts to determine if more exposure to drawing informal inferences with the sampling distribution before teaching the procedures of formal statistical inference would support students' transition to formal statistical inference would also benefit the statistics education community.

Final Remarks

With this study, I sought to add to the knowledge base of the development of informal inferential reasoning and its connection to student's formal inferential reasoning. For the majority of introductory statistics students in this study, their informal inferential reasoning did develop due to their regular classroom instruction. Expanding upon this development with informal statistical inference tasks such as those completed by the interviewees in this study may

help students to bridge the gap between descriptive and inferential statistics. There was some evidence of this as the interviewees were stronger formal inferential reasoners than the other introductory statistics students in the study.

Students demonstrating above average informal inferential reasoning at the beginning of their study of introductory statistics posted significant gains in their informal inferential reasoning as did other students in this study. At the end of the study, they were also stronger formal inferential reasoners compared to the other students. In addition, the strongest relationship between their informal and formal inferential reasoning occurred between their informal inferential reasoning related to the sampling distribution and their formal inferential reasoning. With the sampling distribution as the key to formal statistical inference, this is indeed a desirable outcome. This points to the benefit of the development of informal inferential reasoning in the K-12 curriculum to facilitate a successful transition to formal statistical inference.

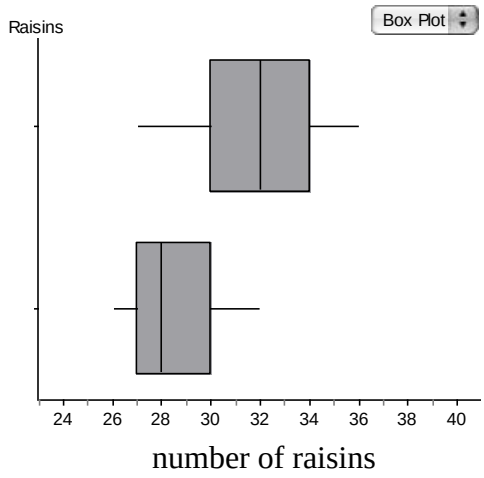
Following the interviewees' informal inferential reasoning through their study of introductory statistics suggests that delaying the use of the formal procedures of statistical inference would give students the opportunity to grasp the power of the sampling distribution. Prior to learning the formal procedures for constructing confidence intervals and conducting hypothesis tests, students need to fully comprehend the keys to formal inference that exist with the sampling distribution. Without this, students will only have their procedural knowledge of formal inference to guide them. Proper interpretation of any statistical analysis requires a comprehensive understanding of the underlying features of formal statistical inference.

Appendix A

Classroom Activities

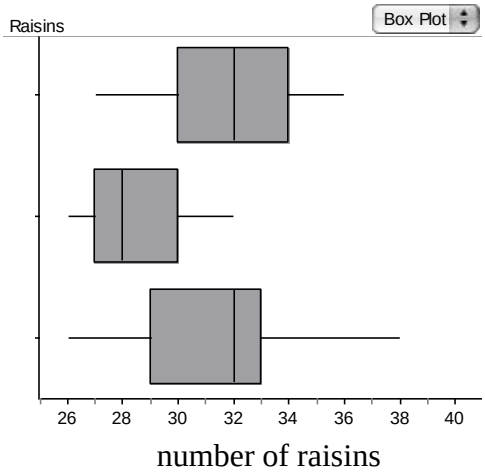
Comparing Distributions Activity **Name_____**

1. Examine the following box plots to help compare the number of raisins per box for two different brands. Sixty ½ ounce boxes of each brand of raisins were examined.

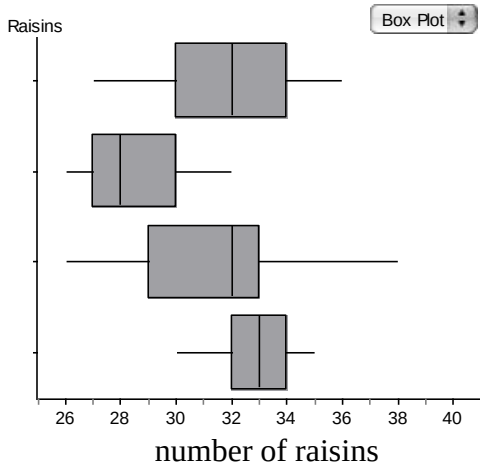


You have a friend who loves raisins. Which brand of raisins would you recommend he/she buy? How would you convince him/her to buy that brand?

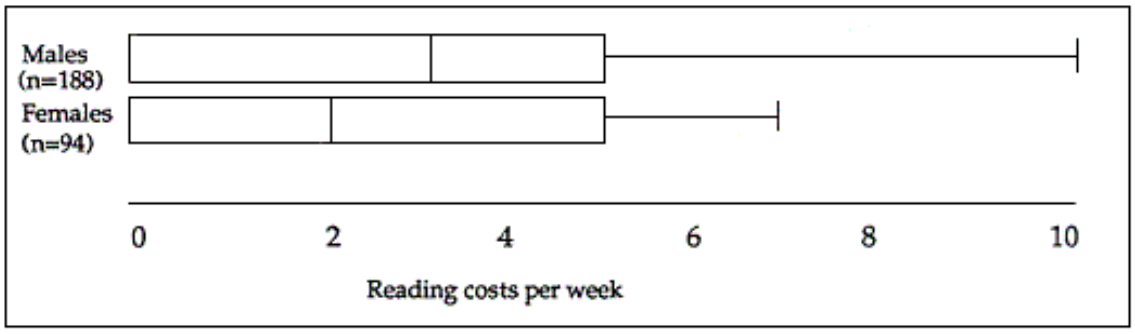
2. Would your recommendation be the same if there were three brands of raisins?



3. What if there were four brands of raisins?



4. Stephen wants to investigate the spending habits of males and females. He compares the amounts spent per week on reading materials by males and females in a random sample of college students by generating the following plots.



What can Stephen conclude?

5. When comparing these distributions, what did you think were the most important aspects to consider?

Sampling to Estimate Probability Activity **Name_____**

Tossing Pigs Activity

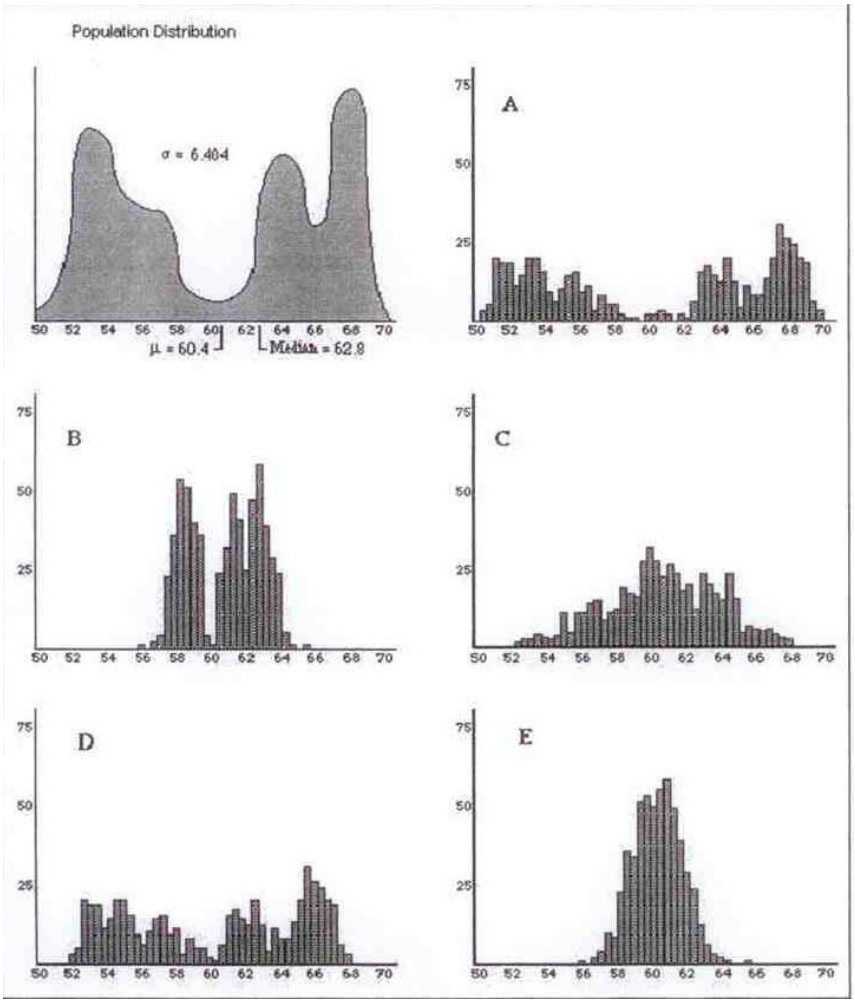
Tossing pigs from the Pass the Pigs game

With your group members, predict the probability that the pig lands on its back when it is tossed. Record the details of your investigation and be prepared to support your prediction with those details.

Sampling Distribution Activity **Name**_____

PART 1

The distribution for a population of test scores is displayed below on the left. Each of the other five graphs labeled A through E represents possible distributions of sample means for random samples drawn from the population.



1. Which graph represents a distribution of sample means for 500 samples of size 4? (circle one)

A B C D E

2. I expect this sampling distribution to have (circle one) **less, the same, more** variability than the population?

3. Which graph represents a distribution of sample means for 500 samples of size 16? (circle one) A B C D E

4. I expect this sampling distribution to have (circle one) **less, the same, more** variability than the first sampling distribution?

PART 2

1. The student body at many community colleges is considered a commuter population. The following question was asked of the Student Affairs Office: "How far (one way) does the average community college student commute daily?" The office answered: "Approximately 10 miles." Sam, a student, believed it was more than 10 miles and decided to test the statement. He took a sample of 50 students. The population standard deviation (σ) is known to be 5 miles.

a. What is the unknown population parameter? _____

b. Construct the Sampling Distribution for samples of 50 students' commuting distances.

c. Suppose Sam found the average commuting distance, $\bar{X}$, of his sample of 50 students to be 10.75 miles. What do you think about the original statement of 10 miles?

2. A psychologist wants to determine whether the average time it takes an adult to react to a certain emergency situation is really 0.56 seconds as claimed by others. From similar studies, she can assume the standard deviation is 0.1 seconds. She decides to use a random sample of 35 adults.

a. What is the unknown population parameter? _____

b. Construct the Sampling Distribution for samples of 35 adult reaction times.

c. Suppose the psychologist found the average reaction time, $\bar{X}$, of her sample of 35 adults to be 0.61 seconds. What do you think about the claim of 0.56 seconds?

Appendix B

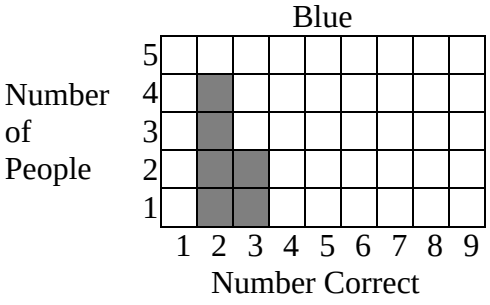
First Task-Based Interview: Comparing Distributions

Two schools are comparing some classes to see which is better at quick recall of 9 math facts. In each part of this question you will be asked to compare different classes. Each box is one person's test score.

PART 1

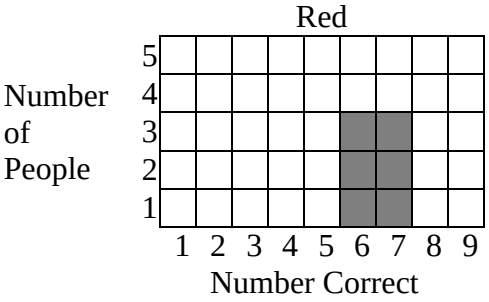
Scores for Blue Class:

2, 2, 2, 2, 3, 3



Scores for Red Class:

6, 6, 6, 7, 7, 7

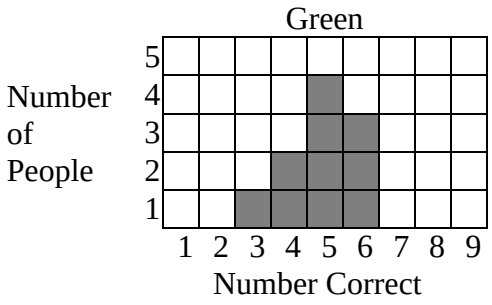


Did the two classes score equally well or did one of the classes score better?

PART 2

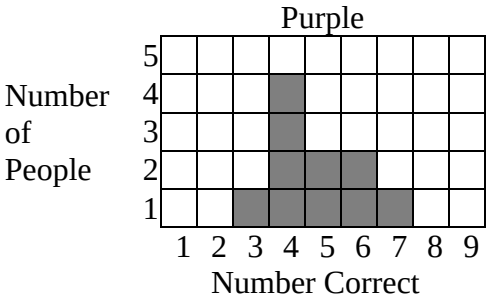
Scores for Green Class:

3, 4, 4, 5, 5, 5, 5, 6, 6, 6



Scores for Purple Class:

3, 4, 4, 4, 4, 5, 5, 6, 6, 7

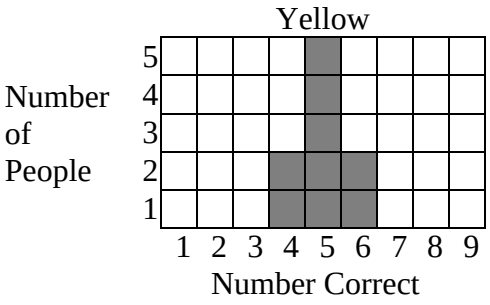


Did the two classes score equally well or did one of the classes score better?

PART 3

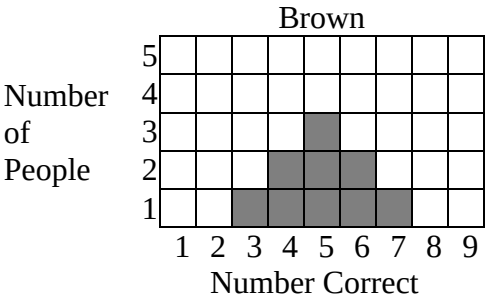
Scores for Yellow Class:

4, 4, 5, 5, 5, 5, 5, 6, 6



Scores for Brown Class:

3, 4, 4, 5, 5, 5, 6, 6, 7



Did the two classes score equally well or did one of the classes score better?

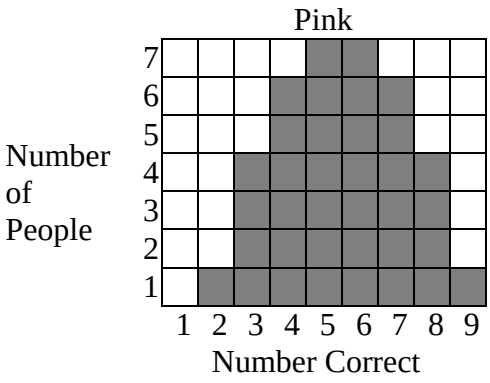
PART 4

Scores for Pink Class:

2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5,

5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8,

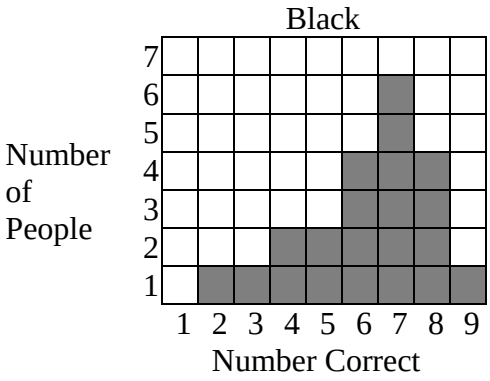
8, 8, 8, 9



Scores for Black Class:

2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,

8, 8, 8, 8, 9

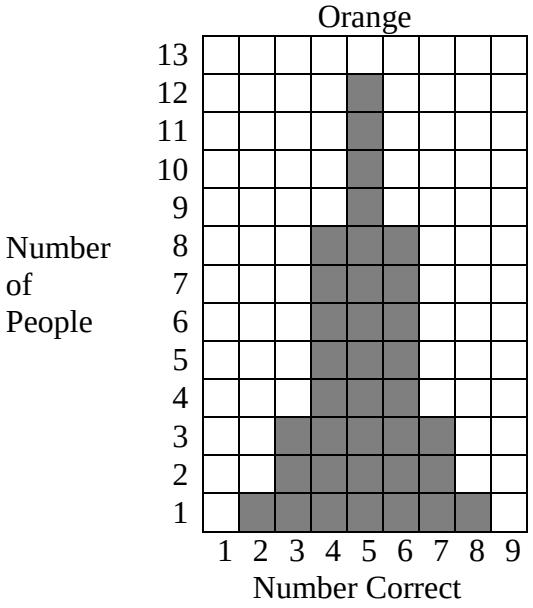


Did the two classes score equally well or did one of the classes score better?

PART 5

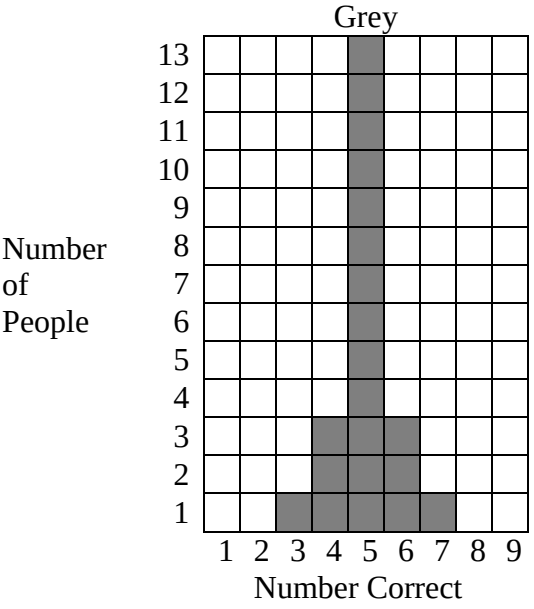
Scores for Orange Class:

2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 8



Scores for Grey Class:

3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7



Did the two classes score equally well or did one of the classes score better?

Protocol for the First Task-Based Interview: Comparing Distributions

Classroom Activity In the Comparing Distributions Activity, you wrote that the most important aspects to consider were_____. Can you talk about that a little more?

PART 1 After students answer the question:

> How would you convince someone else of this?

PART 2 After students answer the question:

> How would you convince someone else of this?

If students answer that the Green class scored better:

> What if someone else said that the Purple class scored better because one student scored a 7 and no one scored a 7 in the Green class? How would you convince this person that the Green class scored better?

If students answer that the Purple class scored better:

> What if someone else said that the Green class scored better because more students scored 5 and 6 in the Green class? How would you convince this person that the Purple class scored better?

PART 3 If students answer that the classes scored equally well or that the Yellow class scored better:

> What if someone else said that the Brown class scored better because one student scored a 7 and no one scored a 7 in the Yellow class? How would you argue that?

If students answer that the Brown class scored better:

> What if someone else said that the Yellow class scored better because more students scored a 5 than in the Brown class? How would you convince this person that the Brown class scored better?

PART 4 If students answer that the classes scored equally well:

> What if someone else said that the Pink class scored better because more students scored 5 and 6 than in the Black class? How would you argue that?

> What if someone else said that the Black class scored better because most of their students scored a 7? How would you argue that?

If students answer that the Black class scored better:

> What if someone else said that the Pink class scored better because more students scored 5 and 6 than in the Black class? How would you argue that?

If students answer that the Pink class scored better:

> What if someone else said that the Black class scored better because most of their students scored a 7? How would you convince this person that the Pink class scored better?

PART 5 If students answer that the classes scored equally well:

> What if someone else said that the Orange class scored better because one student scored an 8? or because more students scored 4 and 6 than in the Grey class? How would you argue that?

> What if someone else said that the Grey class scored better because more students scored a 5 than in the Orange class? How would you argue that?

If students answer that the Orange class scored better:

> What if someone else said that the Grey class scored better because more students scored a 5 than in the Orange class? How would you argue that?

If students answer that the Grey class scored better:

> What if someone else said that the Orange class scored better because one student scored an 8? or because more students scored 4 and 6 than in the Grey class? How would you argue that?

Appendix C

Second Task-Based Interview: Estimating a Probability

PART 1

Tossing Monopoly Houses Activity

Here are my results from 1,000 tosses of the Monopoly house.

House landed	Number
Upright	176
On roof	525
On side	299

PART 2

Green Beads Activity

This jar contains approximately 2000 colored beads. Could you give an estimate of the number of green beads in the jar? How would this be reported as a proportion?

Protocol for Second Task-Based Interview: Estimating a Probability

Classroom Activity In the Tossing Pigs Activity, you predicted the probability that the pig would land on its back to be_____. Can you tell me how you decided on that?

PART 1

a. If you were to toss this Monopoly house, how do you think it would most likely land?

>Why do you think it is most likely to land that way?

> How would you convince someone that this is the side that is most likely to land that way?

b. What do you think the probability is that the house will land upright?

>How did you decide on that probability?

c. Could you find the probability that the Monopoly house will land upright?

> Do you have a strategy in mind for how you will find the probability?
> How would you convince another classmate that this is the probability?

d. Here are my results from 1,000 tosses of the Monopoly house. Would that help you?

If students do not use the results from 1,000 tosses:

> What if a classmate said the probability was .176 based on these results? How would you respond to that?

PART 2

Before students begin:

>What are some of the ways you could go about estimating this number?

Once students begin to take samples with the slotted paddle:

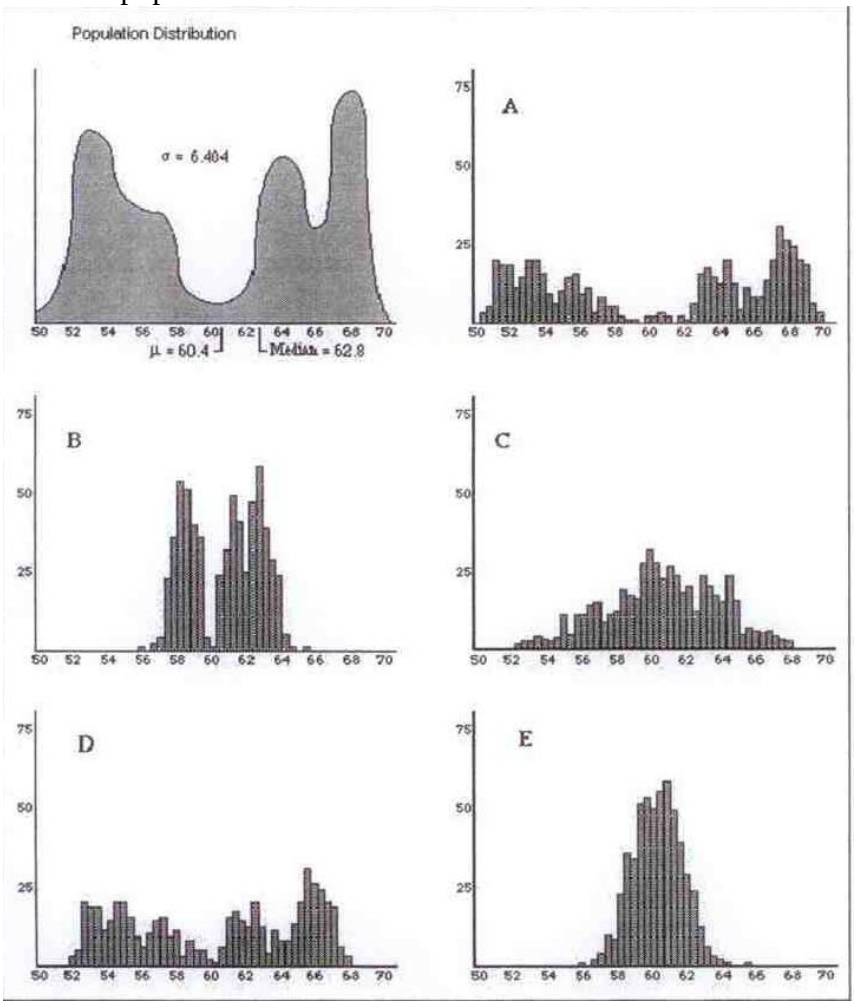
> Do you have a strategy in mind for how you will estimate the number of green beads?

Appendix D

Third Task-Based Interview: Sampling Distribution

PART 1

The distribution for a population of test scores is displayed below on the left. Each of the other five graphs labeled A through E represents possible distributions of sample means for random samples drawn from the population.



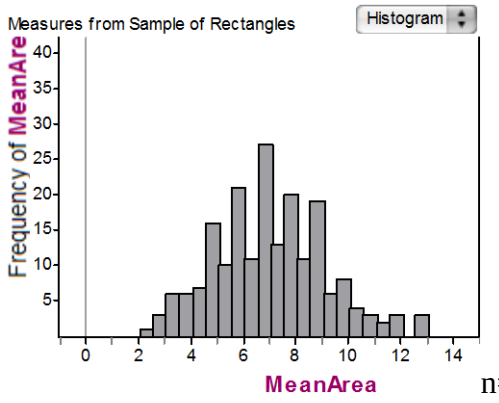
1. Which graph represents a distribution of sample means for 500 samples of size 4? (circle one)
A B C D E

2. I expect this sampling distribution to have (circle one) **less, the same, more** variability than the population?

3. Which graph represents a distribution of sample means for 500 samples of size 16? (circle one) A B C D E

4. I expect this sampling distribution to have (circle one) **less, the same, more** variability than the first sampling distribution?

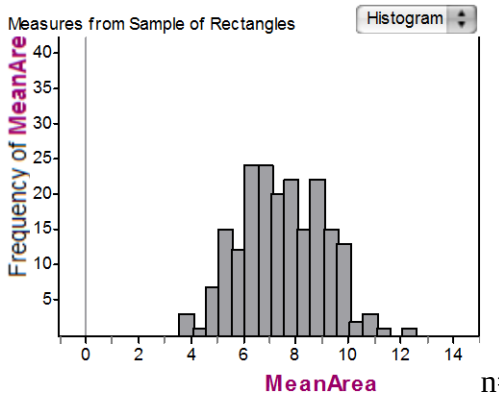
PART 2



n=5

(1) Approximately what values of the sample mean for samples of size 5 would be reasonably likely?

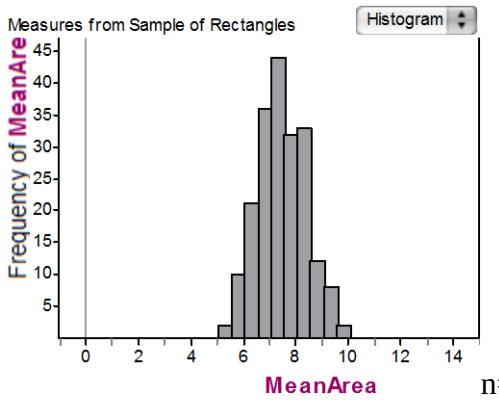
(2) Rare events are defined as those that will occur less than 5% of the time. What values of the sample mean for samples of size 5 would you consider rare?



n=10

(1) Approximately what values of the sample mean for samples of size 10 would be reasonably likely?

(2) What values of the sample mean for samples of size 10 would you consider rare?



n=25

(1) Approximately what values of the sample mean for samples of size 25 would be reasonably likely?

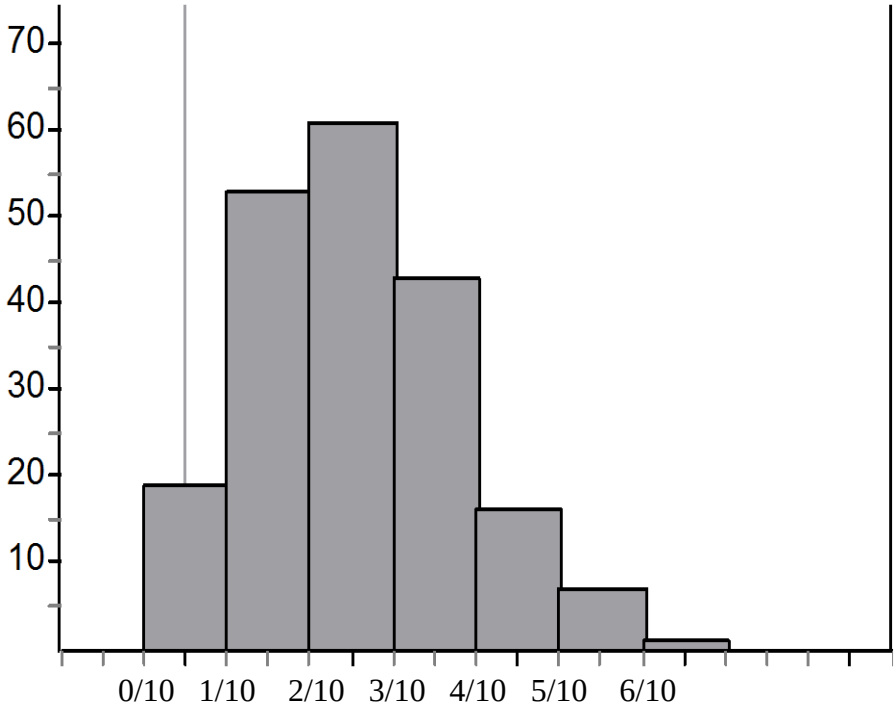
(2) What values of the sample mean for samples of size 25 would you consider rare?

PART 3

Return to the Monopoly Houses Activity

You estimated the probability that a house would land upright to be _____.

Below is the distribution for 200 samples of size 10 for the proportion of houses that landed upright.



Can you determine if the probability that a hotel will land upright is the same as that for a house?

Protocol for Third Task-Based Interview: Sampling Distribution

Classroom Activity

In the Sampling Distribution Activity, you wrote that you thought the original statement of 10 miles was_____. Can you talk about that a little more?

You also wrote that the claim of 0.56 seconds was _____. Can you talk about that a little more?

PART 1

Questions 1 and 2:

>How confident are you that you chose the correct graph?

>Can you talk about why you expected this sampling distribution to have less, the same, or more variability than the population?

Questions 3 and 4:

>How confident are you that you chose the correct graph?

> Can you talk about why you expected this sampling distribution to have less, the same, or more variability than the population?

PART 2

Random Rectangles in Fathom

Show students how the sampling distribution of mean areas is generated from samples of rectangles and their areas. Explore what happens to the sampling distribution of 200 sample means when the sample size changes from 5 to 10 to 25.

>Is the sampling distribution affected by the change in sample size?

>If yes, how so?

>Is there an aspect of the sampling distribution that seems unaffected by the change in sample size?

> How would you explain these changes? Or the aspects that do not change?

Examine the sampling distribution graphs of the sample sizes of 5, 10, and 25 and briefly discuss mean areas that would be likely or expected and those that would be unlikely or rare.

>What about these sampling distributions indicate what mean areas you would consider likely and those you would consider rare?

PART 3 After introducing the sampling distribution graph:

>Is there a way you could describe how this graph might have been created to someone else?

Before answering the question about the hotels:

> I have Monopoly hotels that resemble the houses. Do you think the hotels will behave the same as the houses?

>How do you think the distribution for 200 samples of size 10 might look for the hotels that land upright?

Question about hotels:

> Can you determine if the probability that a hotel will land upright is the same as that for a house?

Once students answer the question about the hotels:

>Can you explain how you determined that these probabilities are the same (or different)?

>What convinces you that the probability that a hotel will land upright is the same (or different) than the probability that a house will land upright?

If students have difficulty getting started:

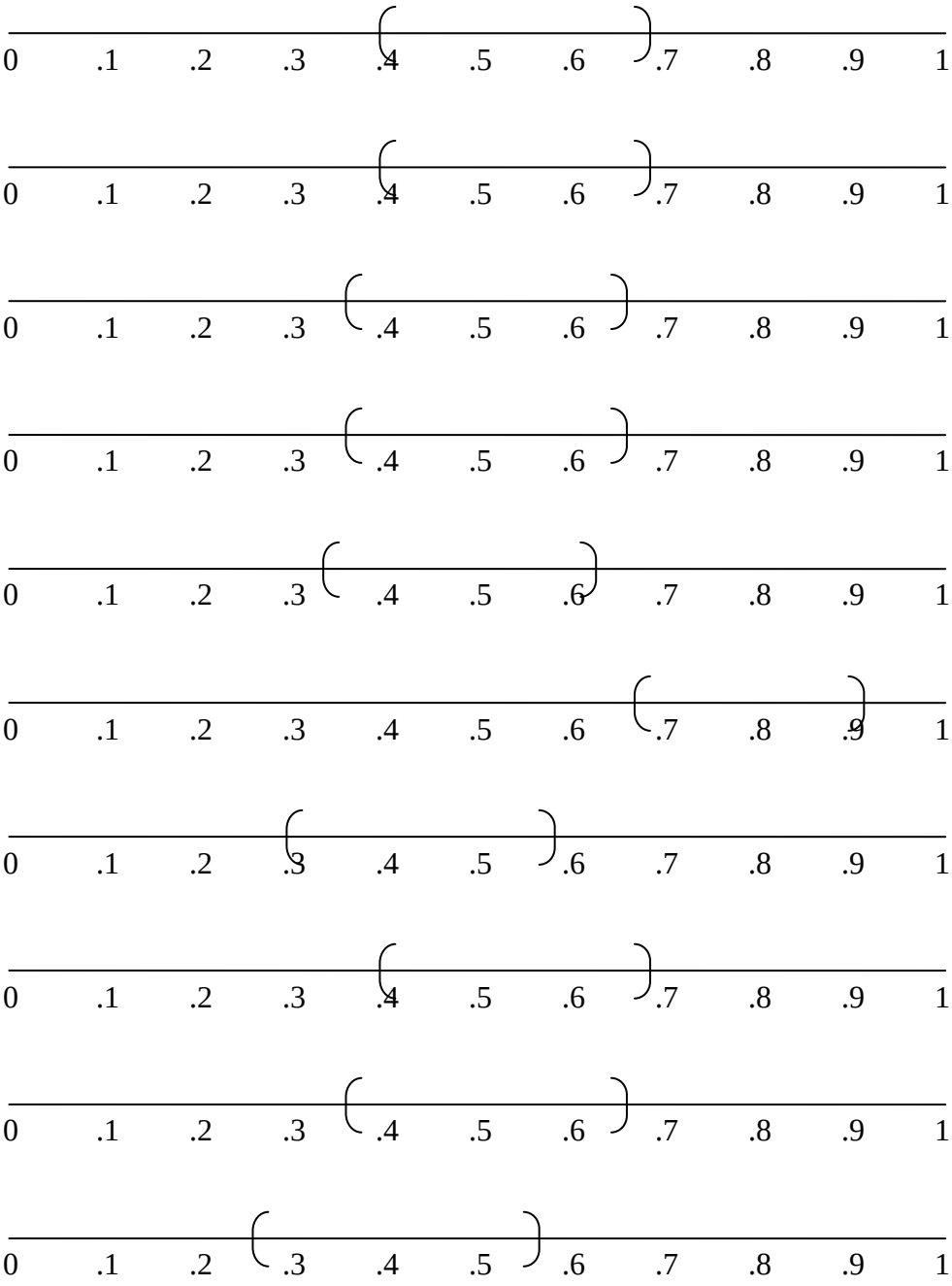
>How could this distribution of 200 sample proportions help you to determine if the probabilities are the same?

>How did you determine this probability for the Monopoly houses?

Appendix E

Fourth Task-Based Interview: Formal Statistical Inference

PART 1 Below are ten 90% confidence intervals for the proportion of red beads in the container based on class results from 10 random samples.



In our second meeting, you estimated the proportion of green beads in the container to be _____.

Based on these intervals, can you say anything about the proportion of red beads?

PART 2

1. Draw a sample from the jar of 2000 beads. Estimate the proportion of red beads in the population with a 90% confidence interval.

2. Interpret this interval.

PART 3

A student believes the proportion of red beads in the container is 70%. Use your sample from Part 2 to test this hypothesis. Would you agree?

Protocol for Fourth Task-Based Interview: Formal Statistical Inference

PART 1 If students have difficulties answering the question about these confidence intervals:

>What makes these confidence intervals different from one another?

>Are there similarities between these confidence intervals?

>What does it mean to be "90% confident"?

>What does this tell you about the proportion of red beads in the container?

PART 2 If students have difficulty getting started:

>How could your sample help in constructing the confidence interval?

>What are the components of a confidence interval?

If students have difficulty with the formula, it will be provided:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

>Can you tell me what $\hat{p}$ represents?

>Can you tell me what $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ represents?

Once students have constructed the confidence interval:

>If another classmate did not know what this confidence interval meant, how would you explain what it means to him/her?

PART 3 If students have difficulty getting started:

>How can you use your result in Part 2?

>Would constructing the sampling distribution be helpful?

If students have difficulty with the formula for the test statistic, it will be provided: $z = \dfrac{\hat{p} - p_o}{\sqrt{\dfrac{p_o(1-p_o)}{n}}}$

>What is a z-value?

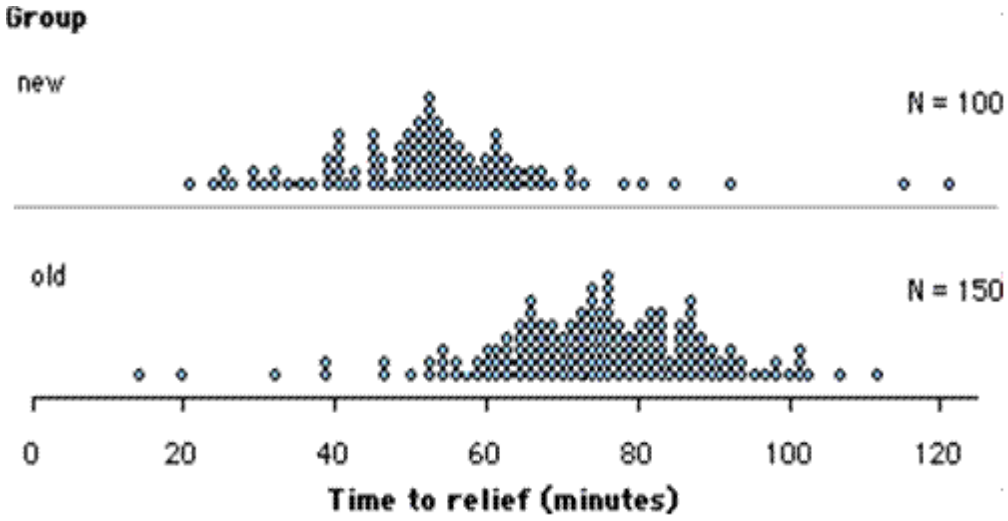
>Can you tell me what $\hat{p}$ - p_o represents?

Appendix F

Pre/Posttest Assessment

Pretest **Name_____**

A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, 250 people were randomly selected from a larger population of patients with headaches. 100 of these people were randomly assigned to receive the new formula medication when they had a headache, and the other 150 people received the old formula medication. The time it took, in minutes, for each patient to no longer have a headache was recorded. The results from both of these clinical trials are shown below.



The items below present statements made by two different statistics students. For each statement, indicate whether you think the student's conclusion is valid.

1. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

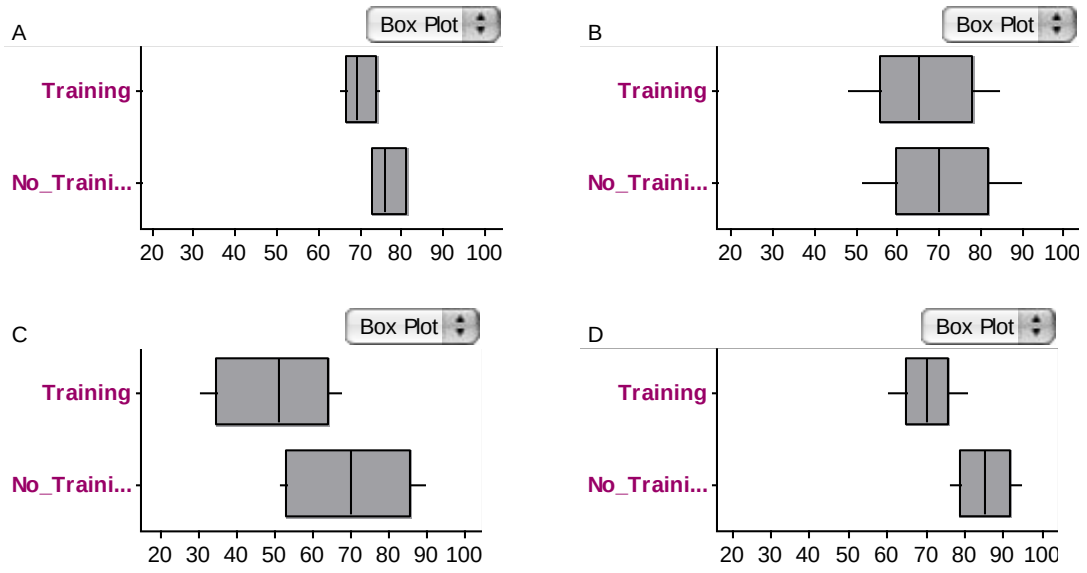
a. Valid b. Not valid

2. The average time for the new formula to relieve a headache is lower than the average time for the old formula. I would conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.

a. Valid b. Not valid

Suppose that there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in their height, weight, and strength. They are randomly assigned to one of two groups. One group gets an additional weight-training program. The other group gets the regular training program without weights. All the students from both groups run the race and their times are recorded, so that the data could be used to compare the effectiveness of the two training programs.

Presented below are some possible graphs that show boxplots for different scenarios, where the running times are compared for the students in the two different training programs (one with weight training and one with no weight training). Examine each pair of graphs and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different training programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment.)



3. Which set of boxplots show the MOST convincing evidence that the weight-training program was more effective in DECREASING athletes' running times?

A B C D

4. Which set of boxplots shows the LEAST convincing evidence that the weight-training program was more effective?

A B C D

A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. They want to estimate the probabilities.

5. Which of the following methods is most appropriate?

a. Since there are four possible outcomes, assign a probability of 1/4 to each outcome.

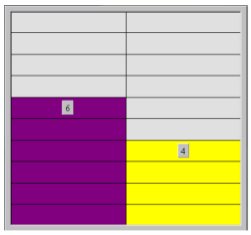
b. Toss the plastic dog many times and see what percent of the time each outcome occurs.

c. Simulate the data using a model that has four equally likely outcomes.

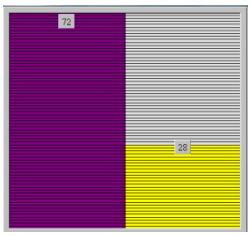
d. None of the above.

Below are the results of 10, 100, and 1,000 spins of the same spinner.

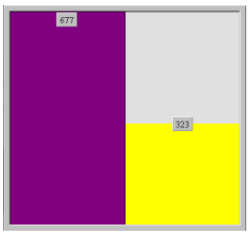
10 spins



100 spins



1,000 spins



6. What is the best approximation of the probability of getting yellow when using this spinner?

a. 0.4 b. 0.28 c. 0.323

A certain manufacturer claims that they produce 50% brown candies. Sam buys a large family size bag of these candies and Kerry buys a small fun size bag.

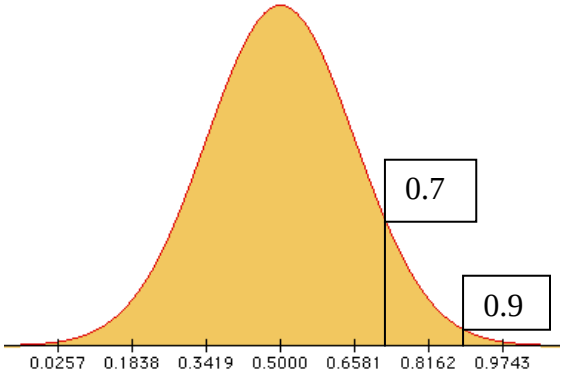
7. Sam discovers that his bag contains 20% brown candies. Is this surprising?

Yes No

8. Kerry discovers that her bag contains 80% brown candies. Is this surprising?

Yes No

Below is a model for the sampling distribution of the proportion of heads you would expect when a fair coin is balanced 10 times on its edge. The model was created using simulation for the "fair" process of obtaining heads 50% of the time.



Two people balance a coin 10 times. One person obtains a proportion of heads of 0.7 and the other 0.9. These results are marked on the distribution.

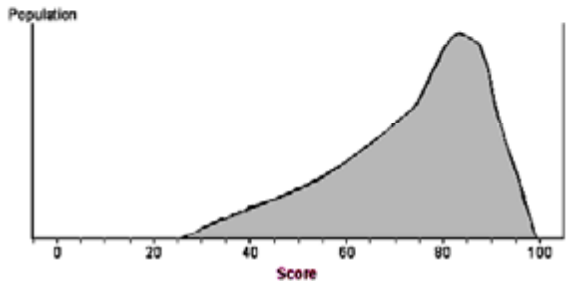
9. Is it reasonable to conclude that the coin is "fair" with a sample proportion of 0.7 heads?

Yes No

10. Is it reasonable to conclude that the coin is "unfair" with a sample proportion of 0.9 heads?

Yes No

Shown below is a graph of scores from many sections of students who have taken a particular exam. For this population, the average score is 74.



A random sample of 50 students in the class this year, given the exact same exam, had a mean exam score of 78.

11. Do you think that the teacher can say that this year's students did better on average than what would be expected?

Yes No

12. Do you think this higher sample average score could just be due to chance?

Yes No

Additional Questions on the Posttest

A high school statistics class wants to estimate the average number of chocolate chips in a generic brand of chocolate chip cookies. They collect a random sample of cookies, count the chips in each cookie, and calculate a 95% confidence interval for the average number of chips per cookie (18.6 to 21.3). Indicate if each interpretation is valid or invalid.

13. We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips.

a. Valid b. Invalid c. Unsure

14. We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips.

a. Valid b. Invalid c. Unsure

15. We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie.

a. Valid b. Invalid c. Unsure

16. Two different samples will be taken from the same population of test scores where the population mean and standard deviation are unknown. The first sample will have 25 data values, and the second sample will have 64 data values. A 95% confidence interval will be constructed for each sample to estimate the population mean. Which confidence interval would you expect to have greater precision (a smaller width) for estimating the population mean?

a. I expect the confidence interval based on the sample of 64 data values to be more precise.

b. I expect both confidence intervals to have the same precision.

c. I expect the confidence interval based on the sample of 25 data values to be more precise.

Each of the 110 students in a statistics class selects a different random sample of 35 Quiz scores from a population of 5000 scores they are given. Using their data, each student constructs a 90% confidence interval for μ the average Quiz score of the 5000 students.

17. Which of the following conclusions is correct?

a. About 10% of the sample means will not be included in the confidence intervals.

b. About 90% of the confidence intervals will contain μ.

c. It is probable that 90% of the confidence intervals will be identical.

d. About 10% of the raw scores in the samples will not be found in these confidence intervals.

A research article reports the results of a new drug test. The drug is to be used to decrease vision loss in people with Macular Degeneration. The article gives a *p*-value of .04 in the analysis section. Indicate if each interpretation is valid or invalid.

18. The *p*-value of .04 is the probability of getting results as extreme as or more extreme than the ones in this study if the drug is actually not effective.

a. Valid b. Invalid c. Unsure

19. The *p*-value of .04 is the probability that the drug is effective.

a. Valid b. Invalid c. Unsure

20. A research article gives a *p*-value of .001 in the analysis section. Which definition of a *p*-value is the most accurate?

a. the probability that the observed outcome will occur again.

b. the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.

c. the value that an observed outcome must reach in order to be considered significant under the null hypothesis.

d. the probability that the null hypothesis is true.

It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass.

A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches.

21. Which of the following is the most appropriate statistical conclusion?

a. The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.

b. The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.

c. The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.

It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the *p*-value is .0025.

22. Which of the following is the best interpretation of the *p*-value?

a. A *p*-value of .0025 provides strong evidence that Mrs. Rose's class outperformed high school students across the nation.

b. A *p*-value of .0025 indicates that there is a very small chance that Mrs. Rose's class outperformed high school students across the nation.

c. A *p*-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well for this national test.

d. None of the above.

References

American Statistical Association. (2010). *Guidelines for assessment and instruction in statistics education (GAISE) college report*. Retrieved from http://www.amstat.org/education/gaise

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy,reasoning and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42-63. Retrieved from http://www.stat.auckland.ac.nz/~iase /serj/SERJ3(2)_BenZvi.pdf

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland .ac.nz/~iase /publications/17/2D1_BENZ.pdf

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests. *Journal of Statistics Education, 17*(2). Retrieved from http://www.amstat.org/publications/jse/v17n2 /castrosotos.html

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

DelMas, R., Garfield, J., & Chance, B. (1998). Assessing the effects of a computer microworld on statistical reasoning. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W. Wong Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 1083–1089). Nanyang Technological University, Singapore: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/2/Topic8e.pdf

DelMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3). Retrieved from http://www.amstat.org/publications/jse/secure /v7n3 /delmas.cfm

DelMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2) _delMas.pdf

DelMas, R., Ooms, A., Garfield, J., & Chance, B. (2006). Assessing Students' Statistical Reasoning. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-5). Salvador, Bahia, Brazil. International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots7 /6D3_DELM.pdf

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2009). *Stats: Data and models* (3rd ed.). Boston: Pearson Education.

Finzer, W. (2001). *Fathom Dynamic Data Software.* (Version 2.1) [Computer software]. Emeryville, CA: Key Curriculum Press.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York, NY: Springer.

Garfield, J., delMas, R., Chance, B., & Ooms, A. (2006). Assessment resource tools for improving statistical thinking. [Website]. Retrieved from: https://apps3.cehd.umn.edu /artist/index.html

Garfield, J., delMas, R., & Zieffler, A. (2007) Adapting and implementing innovative material in statistics. [Website]. Retrieved from: http://www.tc.umn.edu/~aims/index.htm

Garfield, J., Zieffler, A., & Lane-Getaz, S. (2005). EPSY 3264 Course Packet, University of Minnesota, Minneapolis, MN. Retrieved from: http://www.tc.umn.edu/~aims /aimstopicsComparingGroups.htm

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517–545). Mahwah, NJ: Laurence Erlbaum Associates.

Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics Education in the South Pacific: Vol. 2. Proceedings of the 26th Annual Conference of the Mathematics Education Research Group of Australasia, Auckland* (pp. 366–373). Sydney, NSW: MERGA.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6*(1), 59-98.

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., Finzer, W., Horton, N. J., & Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning, 13*(1), 68-86. doi**:**10.1080/10986065.2011.538299

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic Data Explorations* (Version 1.0) [Computer software]. Emeryville, CA: Key Curriculum Press.

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27-54. Retrieved from http://www.stat.auckland.ac.nz /~iase/serj/SERJ4(1)_Makar_Confrey.pdf

Makar, K., & Rubin, A. (2007). Beyond the bar graph: Teaching informal statistical inference in primary school. In J. Ainley & D. Pratt (Eds.), *Reasoning about Statistical Inference: Innovative Ways of Connecting Chance and Data. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy* (pp. 1-29). The University of Warwick.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105. Retrieved from http://www.stat .auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York: W. H. Freeman and Company.

National Assessment Governing Board. (1994). *Mathematics framework for the 1996 national assessment of educational progress*. Washington, D.C.: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45. Retrieved from http://www.stat.auckland .ac.nz/~iase/serj/SERJ7(2)_ Pfannkuch.pdf

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal, 7*(2), 107-129. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pratt.pdf

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Annual Conference of the International Group for the Psychology of Mathematics Education* (pp. 89–96). Hiroshima, Japan: Hiroshima University.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Shaughnessy, J. M., Canada, D., & Ciancetta, M. (2003). Middle school students' thinking about variability in repeated trials: A cross-task comparison. In N. A. Pateman, B. J. Dougherty, & J. T. Zilliox (Eds.), *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (pp. 159–165). Honolulu, HI: Center for Research and Development Group, University of Hawaii.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51*(3), 225–256.

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal, 7*(2), 59-82. Retrieved from http://www.stat.auckland .ac.nz/serj

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*, 289-312.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistics Review, 67*(3), 223-265.

Zieffler, A., Garfield, J., delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf

VITA

NAME OF AUTHOR: Bridgette Lynn Jacob

DEGREES AWARDED:

Master of Science in Mathematics Education, 2000, Syracuse University

Bachelor of Arts in Mathematics, 1983, Niagara University

PROFESSIONAL EXPERIENCE:

Associate Professor of Mathematics
Onondaga Community College, 2004-present

Mathematics Teacher
Westhill High School, 2000 – 2004

PAPERS AND PRESENTATIONS:

Jacob, B., & Doerr, H. M. (2013). Students' informal inferential reasoning when working with the sampling distribution. In proceedings of the *Eighth Congress of European Research in Mathematics Education*, Antalya, Turkey.

Jacob, B., & Doerr, H. M. (2012). *Navigating the transition from descriptive to inferential statistics through informal statistical inference*. Poster session presented at the 34th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Kalamazoo, MI.

Jacob, B., & Doerr, H. M. (2012). *Examining the links between informal and formal inferential reasoning.* Poster session presented at the research pre-session of the annual meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Doerr, H. M., & Jacob, B. (2011). Investigating secondary teachers' statistical understandings. In proceedings of the *Seventh Congress of the European Society for Research in Mathematics Education*, Rzeszow, Poland.